Computer simulation of demographic forecast

Victoria Yanulevskaya

13.10.2006

University of Joensuu
Department of Computer Science
Master's Thesis

UNIVERSITY OF JOENSUU RESEARCH RELEASE

Population forecast shows how the total population number changes over time and this information is extremely important for planning purposes. National, regional and local planners all need to have some idea of likely future changes in the size and age distribution of the population in their particular areas.

Any population forecast includes uncertainty because human behaviour is quite unpredictable in general. Hence, it is natural to add propagation error as a random component in order to model the uncertainty.

In this thesis, the basic demographic events and measures are first reviewed. These include Lexis diagram, cohort and vital rates. Then the linear grows model and the dynamic stochastic population forecast are introduced and illustrated with numerical examples. Own stochastic forecast of population is constructed and computer implementation details are discussed. Finnish population forecast in 2004-2054 is simulated and the result is compared with United Nation forecast.

Author(s) : Victoria Yanulevskaya

Student number(s): 157755

Title: Computer simulation of demographic forecast

Faculty/Department _Faculty of Science/Computer Science

Nr. of pages: 57. Time: 13.10.2006. Type of work: pro-gradu

Key words: population forecast, linear growth model, uncertainty.

Abstract

Population forecast shows how the total population number changes over time and this information is extremely important for planning purposes. National, regional and local planners all need to have some idea of likely future changes in the size and age distribution of the population in their particular areas.

Any population forecast includes uncertainty because human behaviour is quite unpredictable in general. Hence, it is natural to add propagation error as a random component in order to model the uncertainty.

In this thesis, the basic demographic events and measures are first reviewed. These include Lexis diagram, cohort and vital rates. Then the linear grows model and the dynamic stochastic population forecast are introduced and illustrated with numerical examples. Own stochastic forecast of population is constructed and computer implementation details are discussed. Finnish population forecast in 2004-2054 is simulated and the result is compared with United Nation forecast.

KEY WORDS: population forecast, linear growth model, uncertainty.

Table of Contents

List of Figures

List of Tables

1. Introduction

Forecasting is an important part of decision making and planing. Webster's dictionary defines forecasting as an activity “to calculate or predict some future event or condition, usually as a result of rational study or analysis of pertinent data” [16]. Population forecasting is a basis of any other social prediction. In fact, if it is needed to plan evolution of some products or services making, distribution of natural resources and budget receipts, and any other social processes, the population size and its structure should first be considered. Statistical data regarding population refer to the past, action and policy require knowledge of the future. The cost of an annuity taken into account today depends on the future not on the past mortality [13]. The question of the population forecast continues between past and future.

In general, *demography* is a science about population. According to the United Nations Multilingual Demographic Dictionary [26] demography is a scientific study of human populations, primarily with respect to their size, structure and development. In other words, it concerns with the current size and characteristics of human populations, how they were attained, and how they are changing.

The purposes of the demographic forecasting concerns with the economic decision making, such as pension, public health and education planning. World, national, regional and local forecasts are needed to have some idea of likely future changes in the size of the population in a particular area.

The subject of demographic forecasting is the population size, or changing of the size. The population can increase via births and in-migration, and decrease via deaths and out-migration, see Figure 1.

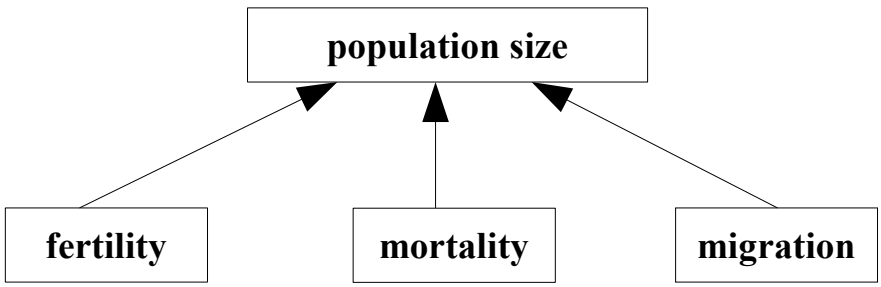

Figure 1. Influences on population size.

Thus, births, deaths and migration form the relevant vital processes. Population forecasts summarize existing information about the likely future development of the vital rates. The mathematical model looks like this:

$$\begin{aligned} &\textit{Population size at year}(t) = \textit{Population size at year}(t-1) + \\ &\textit{Number of newborns} - \textit{Number of deaths} + \\ &\textit{InMigration} - \textit{OutMigration}. \end{aligned} \qquad \textbf{(1)}$$

These data can be obtained through population censuses, special sample surveys, current records of demographic events and so on. This information is usually given by sex and age. Moreover, special demographic measures are also included, for example, *age specific fertility rate* (*ASFR*), *projective mortality rate* (*PMR*), *probability of surviving* and *scales of uncertainty*. Demographic measures characterise different population characteristics, population structure, demographic processes and reproduction of population.

Since future events involve uncertainty, the forecasts are usually not perfect and contain errors. The objective of forecasting is to reduce the forecast error: to produce forecast that are seldom incorrect and that have small forecast error. It can be done by including randomness in the model [15]. C*ohort-component book-keeping* forecast is based on equation (1) and it includes randomness by adding error in a specific form to each component.

The rest of the thesis is structured as follows. In Chapter 2, definitions of demographic events based on Lexis diagram are presented. Chapters 3, 4 and 5 give detailed picture of the main factors that affect population changes: fertility, mortality and migration accordingly. Further linear growth and dynamic stochastic demographic models are given in Chapter 6, where the problem of uncertainty in population forecasts is studied. Chapter 7 is dedicated to extensive description of stochastic population forecast with a toy example model. Computer implementation is covered in Chapter 8 and experimental results are reported in Chapter 9. Concluding remarks are given in Chapter 10.

2. Description of demographic events

Time is the variable of all events in the world, which is the case for demographic processes also. There are two types of demographic time: (1) real calendar time, which is measured by the exact dates of beginning or ending of some event (such as dates of birth and death); and (2) the duration of an event, for example, age since the birth till the observation point. Both types exert influence on person's life hence on demographic events [31]. Probability of some event depends on calendar time (during wars fertility decreases), and on history (old people have higher mortality rate), which is duration dependence. Therefore, one should take into consideration these two times for adequate picture. A way to consider this dual time nature is to use a *Lexis diagram*.

2.1. Lexis diagram

Lexis diagram is one of the main instrument of demographic analysis. It is a rectangular reference grid, where horizontal axis refers to time (t) and the vertical axis to age (x) [28]. For each person, a life line may be drawn that starts at a time and age when the person enters the population and ends at the time when the person exits the population. Typically the entry would occur at birth and exit at death, but entries or exits due to other vital processes (e.g. migration) may occur at other ages.

Demographic events are shown in three-dimensional space on Lexis diagram: *the date of enter* of a given demographic event, *the date of exit,* and *observation period* (if we end observation before the date of exit), time in demographic event at the exit or at the moment of observation [5]. In such a way, we use three coordinates in two-dimensional space to describe any demographic event. Demographic events are birth, death, being *n*-years old, migration, marriage and divorce. Observation period is a time of our interest.

We need to have two coordinates to show any event in the Lexis diagram. The first one is the date of event starting (Time axis) and the second is the age of individual (Age axis). For example, point A in Figure 2 corresponds to an event that starts at exact time 3 when the individual is exactly 4 years old. We can call this event “being exact 4-years old in the 3 year”. Lexis diagram is a line BC for event “being exact 4-

years old", this line is called *line of age* 4.

Exact age is not used often as coordinate in Lexis diagram, usually age rounded to the number of years since birth, the event "being 4-years old in the year 3" corresponds to the line AD in Figure 2.

With Lexis diagram we can localize age and year. If event occurs at age 2, it looks like a horizontal block restricted by *lines of age* 2 and 3, see rectangle E in Figure 2. If we are interested in event, which occurs in year 5, we should consider vertical block restricted by *lines of years* 5 and 6, see rectangle F in Figure 2. Similar to this, the line of age GH represents the event "being during year 5".

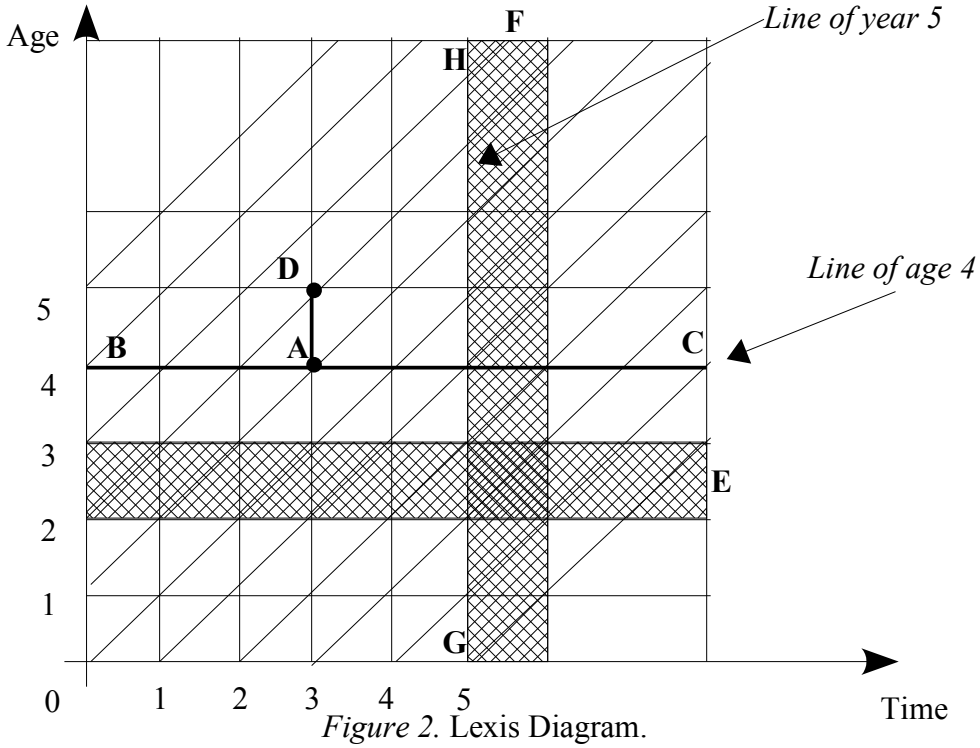

Figure 2. Lexis Diagram.

2.2. Definition of cohort

The third coordinate of Lexis diagram is the date of birth or entering date for example by migration. For simplicity we will use hereinafter only terms "birth" and "death", but remember other alternatives. This is coordinate on horizontal axis (Time axis), but it is presented like an inclined line, which shows the changes of individual's age due to the time, see line AB in Figure 3. It is called *life line*. It has an angle of inclination, which equals to 45 degree. All events that happen with an individual are situated on this line. The date of birth is the intersection of the life line and time axis,

this is the point A in Figure 3.

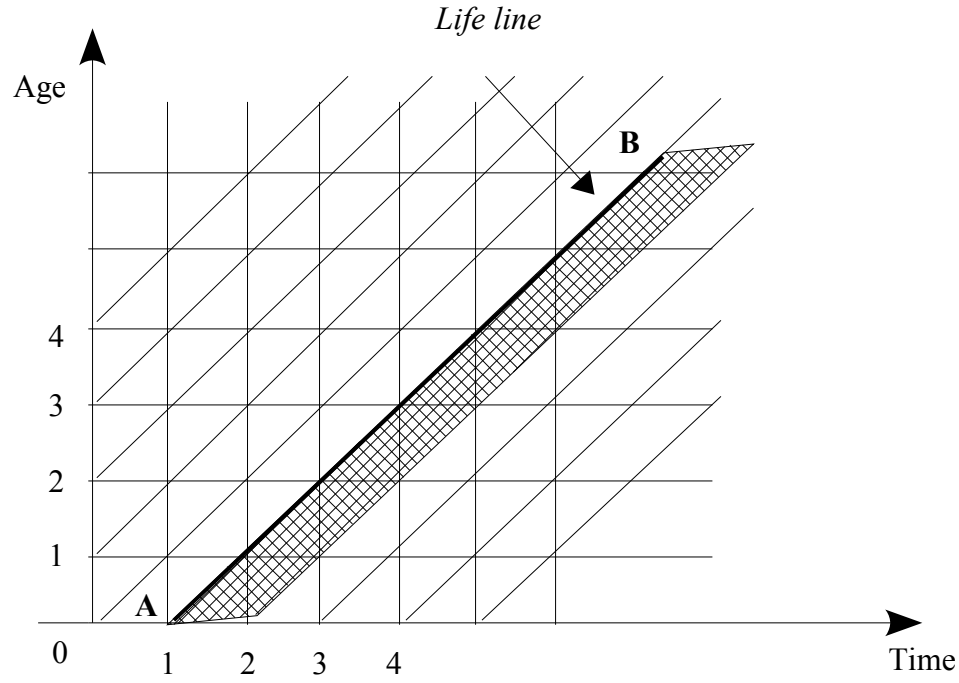

Figure 3. Life line.

The coordinate of event's date of enter and coordinate of individual's birthday are situated on the Time axis, but the first one corresponds to the life line whereas the second one corresponds to the Age axis. Thus, we can use the third coordinate for event's localization. This coordinate is the date of birth and it is represented by the life line. The event A in Figure 4 “being 3-years old” is supplement with “being 3-years old with birthday in year 1”. This event is an intersection of line of age 3 and the life line of the individual with the date of birth in the 1 year.

Three demographic coordinates have strong relation and any of them can be derived from the other two. It is enough to have two coordinates for event's location, the third coordinate can be used for confirmation and for more full description.

In demographic analysis, a *cohort* is defined as a group of people who have had a common experience, gone through the same event during one time period [18]. For example, a group of people, who were born during a particular period or year is called a *birth cohort.* A *marriage cohort*, on the other hand, is a group of people with common date of wedding. Life lines of birth cohort during year 1 are bounded by the lines AB and CD in Figure 5.

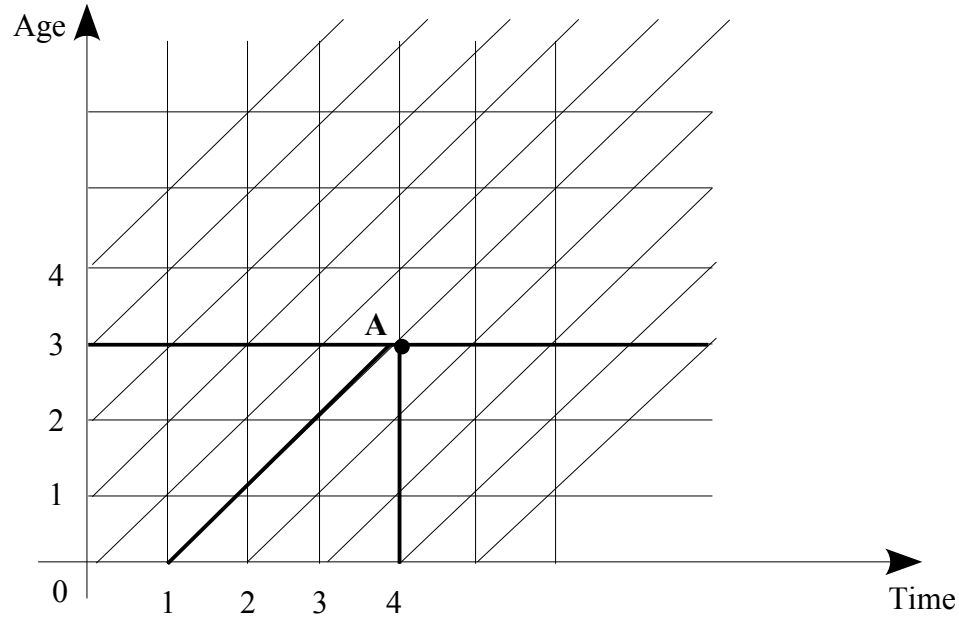

Figure 4. Demographic event is "being 3-years old with birthday in the year 1".

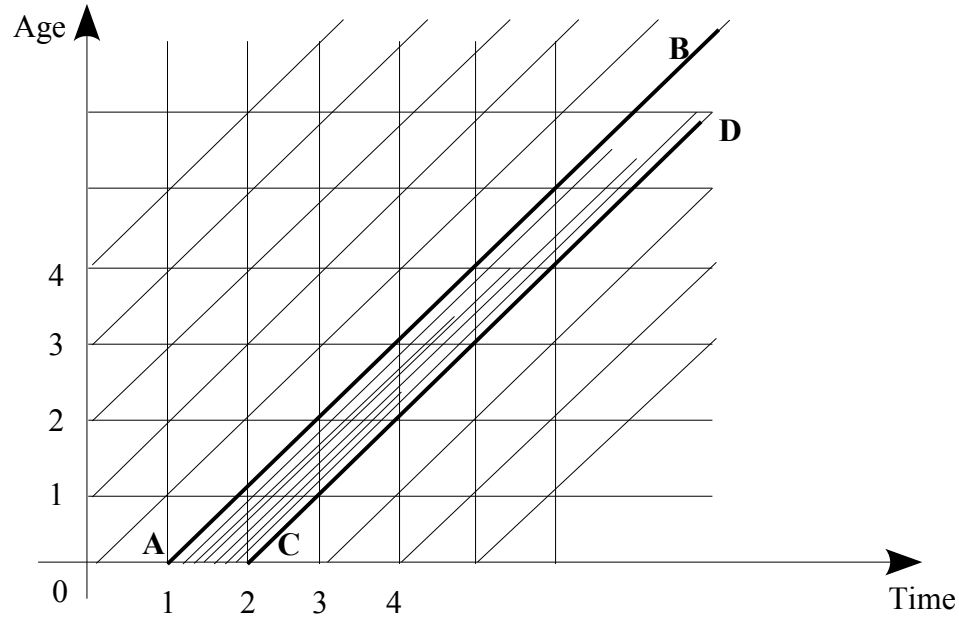

Figure 5. Life lines of a birth cohort.

2.3. Representation of demographic event sets

There are different possibilities to represent demographic events on Lexis diagram. In demographic analysis, there are three *event sets* according to their time orientation.

The first even set is a set of demographic events with the common starting date for

all participants and ending date in a given age. For example, it can be death during age 2 of people who were born in a fixed year 1. This set is an intersection of life lines of cohort for year 1 and the horizontal block restricted by lines of age 2 and 3, see section A in Figure 6. Usually the number of deaths is written near the section. For example, in Finland there were 7 deaths per 100 persons in age 2 for people who were born in 2000 [23].

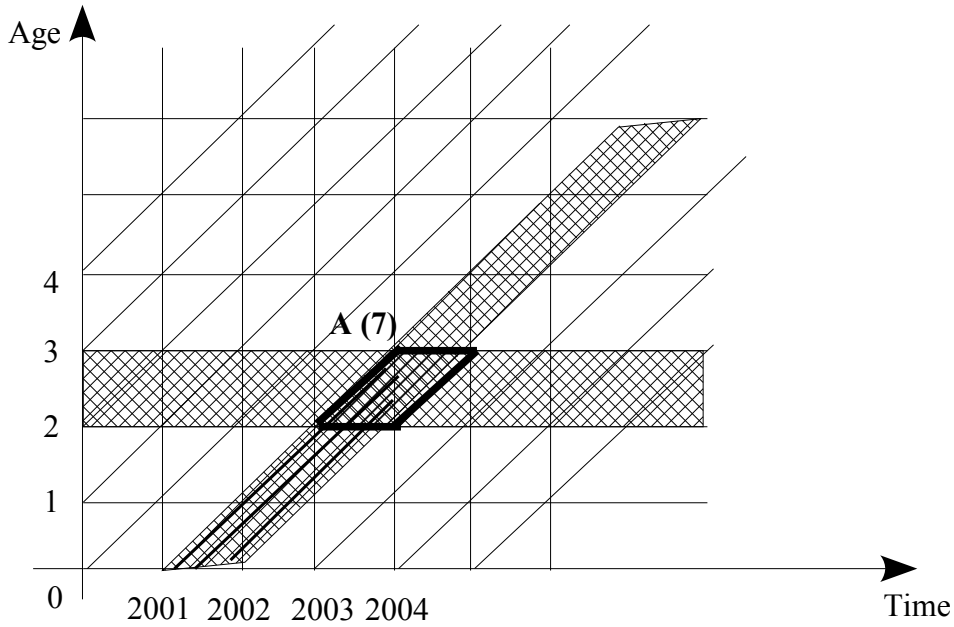

Figure 6. The first type of the set of demographic events.

The second even set is a set of demographic events with the common starting dates and ending point. For example, it can be death during the year 4 for contemporaries. This set is the intersection of the life lines of cohort for year 1 and vertical block restricted by lines of year 4 and 5. See section B in Figure 7. In Finland, there were 6 deaths per 100 persons in 2004 for people who were born in 2000 [23].

The third even set is a set of demographic events with the common ending date during the fixed interval of event's duration. For example, it can be deaths during the year 4 at the age of 2. Here participants are not necessarily contemporaries, they can belong to adjacent cohorts for years 1 and 2, called a *hypothetical cohort*. Hypothetical cohort is an artificial cohort that is built according to the set of age-specific values of vital rates usually for a specific year. In our case, this set is an intersection of horizontal block restricted by the lines of age of 2 and 3, and the vertical block restricted by the lines of year 4 and 5. See section C in Figure 8. In

Finland, there were 10 deaths per 100 persons in 2004 for people who were 2 years old [23].

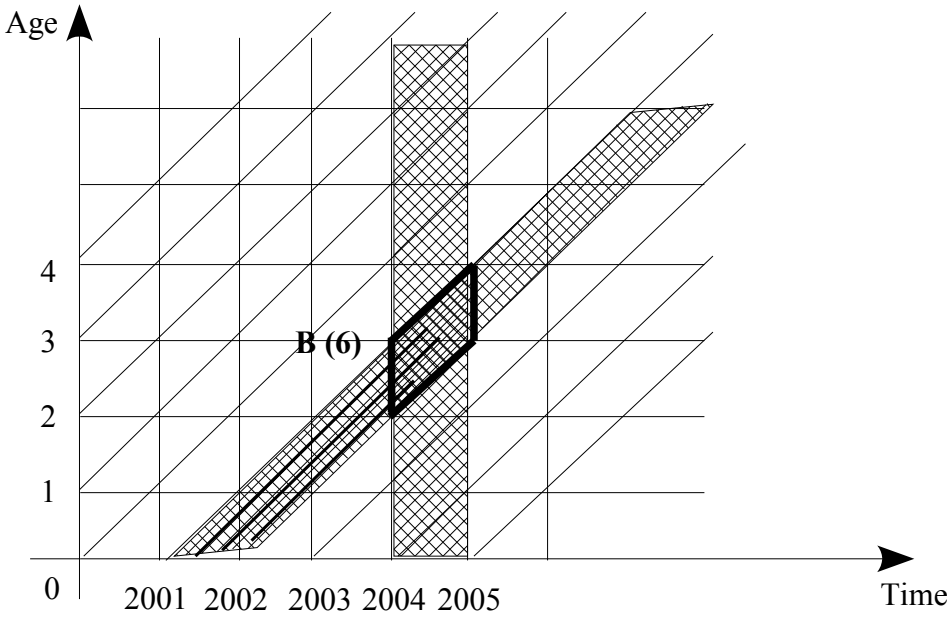

Figure 7. The second type of the set of demographic events.

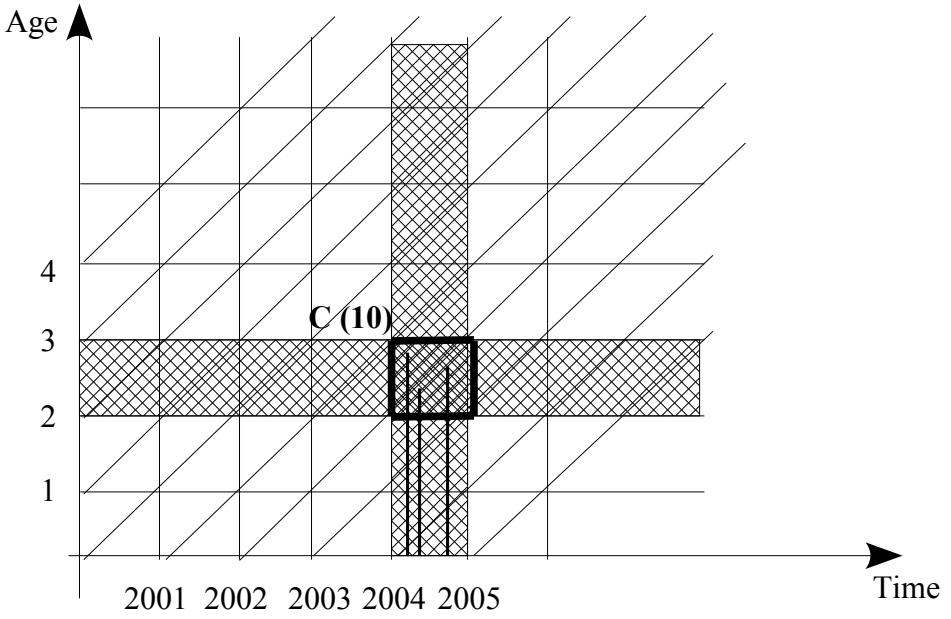

Figure 8. The third type of the set of demographic events.

Three demographic processes influence to future population: *fertility*, *mortality* and *migration*, see equation **(1)**. These will be studied next in the following sections.

3. Fertility

Fertility is the ability to produce children. This term is often used in a comparative sense, with an increase in fertility referring in an increase in the number of children. Fertility rate refers to the actual number of children born alive in a year per 1000 women (of *child bearing age*). There is no uniform definition of the term *live birth* in the world. *World Health Organization* [30] gives the following definition: a live birth of a human being occurs when a foetus is expelled and separated from the mother's body and subsequently shows some sign of life, such as voluntary movement, heartbeat, or pulsation of the umbilical cord, for a brief time. In the absence of such sign, the event is considered a stillbirth. Different countries differ in necessary time after a birth during which infant should have signs of life. It creates difficulties for *international comparisons* [10].

Usually children are born to women in ages 15-45, this interval is called childbearing age, 15 is called the lowest age of childbearing (α) and 45 is the highest age (β). Fertility rate at ages less than 15 is included into the number of births at age 15, and similar fertility at ages over 45 is included into the number of births at age 45 [7].

There are several indexes for fertility measuring, which are used in this work:

- Age specific fertility rate (*ASFR*)
- Total fertility rate (*TFR*)
- Net reproduction rate (*NRR*)
- Fraction of summary fertility
- Mean age of childbearing (*MA*)

Age specific fertility rate is the expected number of births to women in a given age-group per 1000 women in that age-group.

$$ASFR = \frac{{}_nB_m}{{}_nF_m} 1000‰,$$

where ${}_nB_m$ is the number of births to women in the age group [n, m] and ${}_nF_m$ is the number of women in the age group [n, m], where $\alpha \le n < m \le \beta$.

This is the most common way of indicating fertility and 1-year or 5-years age-groups are usually used [20]. This method of measuring fertility removes the distortions produced by variations in the age composition of the population. We can think ${}_nB_m$ as a part of the general number of births, which can be defined in the following way:

$$ {}_nB_m = \Delta_{{}_nB_m} \times B . $$

Similarly, ${}_nF_m$ is a part of women in childbearing age ${}_\alpha F_\beta$, in turn ${}_\alpha F_\beta$ is a part of the total number of people P. Thus, the number of women in an age group $[n, m]$ equals to

$$ {}_nF_m = \Delta_{{}_nF_m} \times {}_\alpha F_\beta = \Delta_{{}_nF_m} \times \Delta_{{}_\alpha F_\beta} \times P. $$

The final expression for the age specific fertility rate:

$$ ASFR = \frac{{}_nB_m}{{}_nF_m} = \frac{\Delta_{{}_nB_m} \times B}{\Delta_{{}_nF_m} \times \Delta_{{}_\alpha F_\beta} \times P} . $$

This representation of *ASFR* shows that these rates avoid influence of the total number of people (since dividing by P), and age structure (since there are age-specific elements in numerator and divider). Figure 9 shows three possibilities to localize ${}_nB_m$.

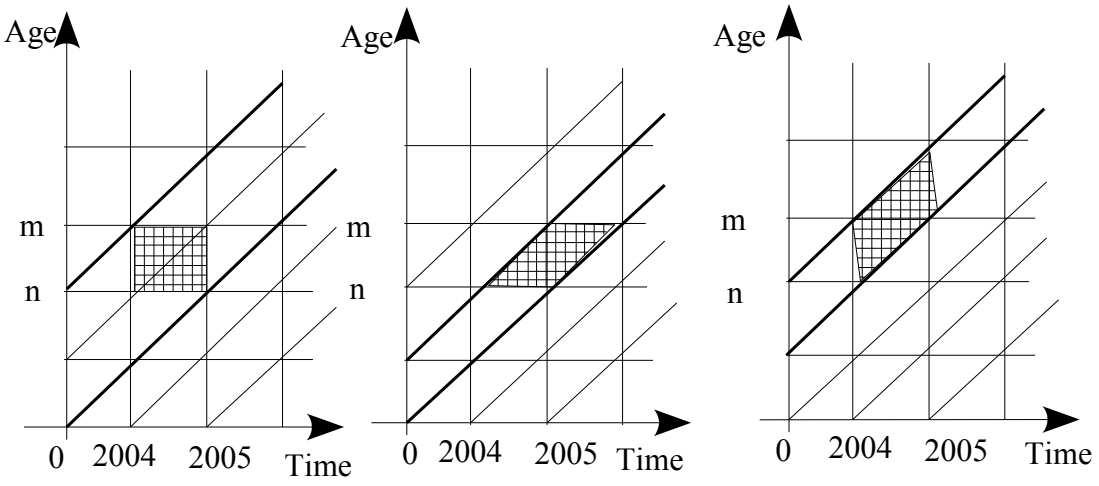

Figure 9. Localization of ${}_nB_m$ at the beginning and at the and of year 2004 (left), ${}_nB_m$ to women that were m years old at the end of 2004 (middle), and ${}_nB_m$ to women that had their m-th birthday during 2004 (right).

We use the first type of localisation. Figure 10 represents *ASFR* in Finland in 2004

[22], and Figure 11 represents part of the Lexis diagram of this data.

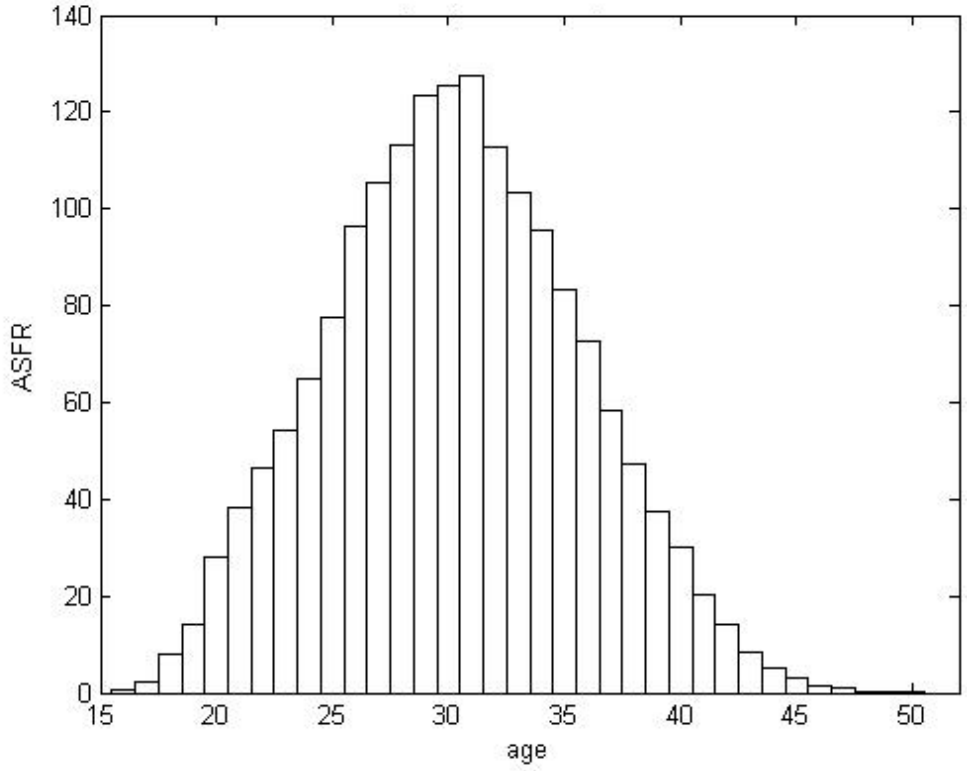

Figure 10. Age-specific fertility rates in Finland in 2004.

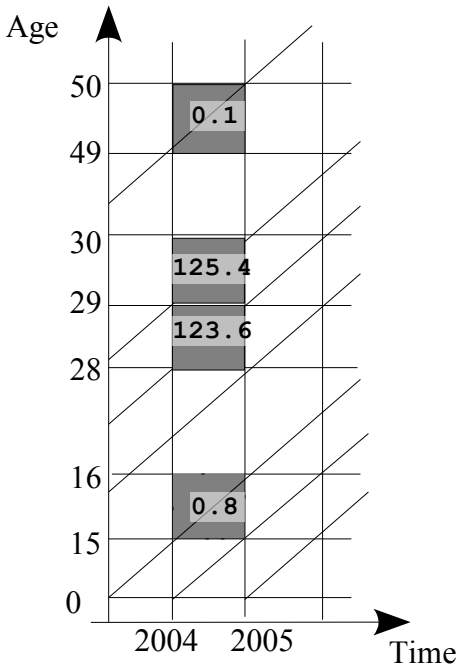

Figure 11. Part of the Lexis Diagram for *ASFR* in Finland in 2004.

Total fertility rate is the average number of live births by a woman during her

lifetime if she survives to the end of her childbearing years under the age-specific fertility rates of a given year. In simpler terms, it is an estimate of the average number of children a woman will have during her childbearing years:

$$TFR = \frac{\sum_{x=\alpha}^{\beta} ASFR_x}{1000},$$

where $ASFR_x$ is the age-specific fertility rate for age x. When 5-year age groups are employed, the *TFR* must be multiplied by 5 since it is the sum of the rates for every individual age:

$$TFR = \frac{5 \times \sum_{x=\alpha}^{\beta} {}_5ASFR_x}{1000}.$$

Thus, the total fertility rate represents the number of children that would be born (ignoring mortality) to a hypothetical group of 1000 women who, as they pass through the reproductive ages, experience the particular age specific birth rates, which the index is based on.

TFR is a more direct measure of the level of fertility than the crude birth rate, since it refers to the number of births per woman. This indicator shows the potential for population growth. High rates will also place some limits on the labour force participation rates for women. Large numbers of children born to women indicate large family sizes that might limit the ability of the families to feed and educate their children. In Finland, *TFR* was 1.73 children born per woman in 2004 [24].

Net reproduction rate is the average number of daughters that would be born to a woman (or a group of women) if she passed through her lifetime conforming to the age-specific fertility and *mortality rates* (see Chapter 4) of a given year. This rate takes into account that some women will die before completing their childbearing years. A value of *NRR*=1 can be interpreted that each generation of mothers is having exactly enough daughters to replace itself in the population. It may be considered as the ratio between the number of females in one generation at a given age and the number of their daughters at the same age, again taking mortality into account.

Fraction of summary fertility is a fraction of *ASFR* in a particular age group [*n*, *m*]:

$$ {}_n\Delta_m = \frac{{}_n ASFR_m}{\sum_n ASFR_m}, $$

where the summation $\sum_n ASFR_m$ is taken over all age groups in childbearing age.

Mean age of childbearing is the average age of mother on childbearing [17]:

$$ MA = \sum_{x=\alpha}^{\beta} (x+0.5)_n \Delta_x . $$

It is related to the population structure. Keyfitz has shown [12] that if the ages of childbearing are spread out, apparently the gain through some children being born earlier more than offsets the loss through those born later. The smaller *MA* values guarantee higher *NRR* values, because more women are alive at mean age of childbearing, thus more children would be born. In Finland, *MA*=29.6 in 2004.

4. Mortality

In general meaning the term mortality relates to death as a component of population change [21]. Mortality rate is the frequency of number of deaths in proportion to the population. It is a mass process of individuals' life stopping. With fertility they form natural population evolution. Again, there are several indexes for mortality rate measuring:

- Age-specific mortality rate (*ASMR*)
- Decline mortality rate (*DMR*)
- Infant mortality rate (*IMR*)
- Survival probability (*P*)

Age-specific mortality rate is the mortality rate limited to a particular age group for males and females separately. The numerator is the number of deaths in that age group, and the denominator is the number of persons in the age group in the population:

$$ASMR = \frac{{}_{n}D_{m}}{{}_{n}P_{m}} 1000‰,$$

where ${}_{n}D_{m}$ is the number of death in age group $[n, m]$, ${}_{n}P_{m}$ is the number of individuals in an age group $[n, m]$, where $x \le n < m \le y$, so that x is the lowest and y is the highest population age. This is the most popular index especially age-specific mortality rate for 5-year groups. As it was shown in Chapter 3, age-specific coefficients allow to avoid influence of the total number of people and age structure. Figure 12 represents logarithm of numbers for males' and females' *ASMR* in Finland during 2004 year.

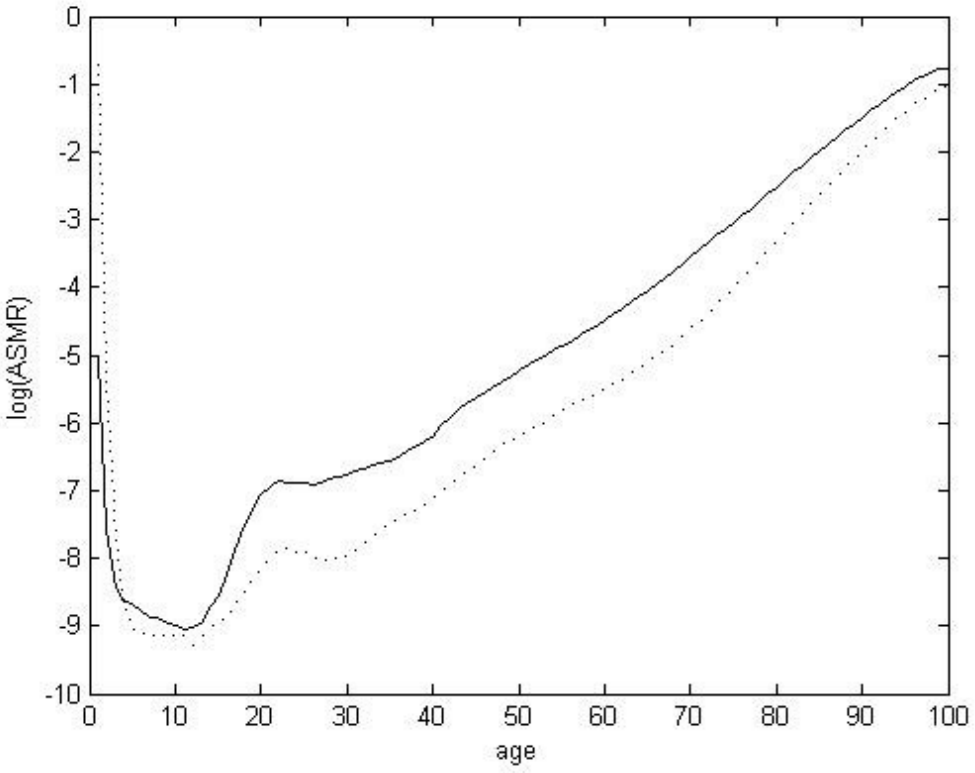

Figure 12. Logarithm of age-specific mortality rates for males (solid) and females (dashed) in Finland in 2004.

Decline mortality rate characterizes the changes of mortality rate and it can be calculated as first time derivative of logarithm of *ASMR*:

$$DMR_x(t) = \frac{\delta}{\delta t} \log ASMR_x(t),$$

where *ASMR's* are taken for some period $t \in [T, T+\Delta]$.

This measure is associated with the total growth of population. Historically, human population has grown very slowly except during the last two centuries. World population now stands at 6 billion, and according to United Nations projections, it will continue to grow through this century [25], see Figure 13.

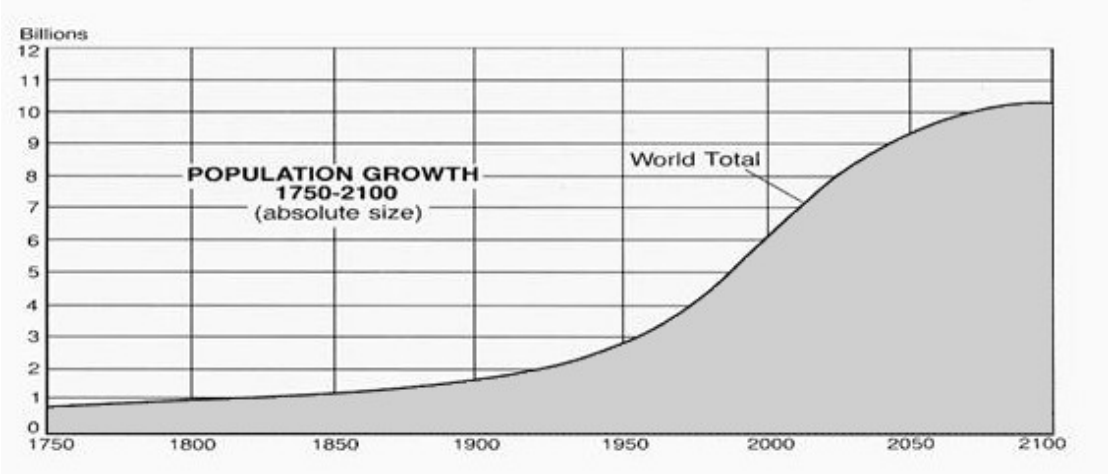

Figure 13. World Population [25].

The cause of the initial mortality decline lies in the development of new medical and public health technologies based on anti-bacterial chemicals and insecticides that reduce disease vectors. Greater declines in the early 20th century attributed to improvements in medical technology, which led to the control of such infectious diseases as tuberculosis, smallpox and cholera. Further improvements in life expectancy are anticipated in most countries [19]. Thus, *DMR* should be taken into account for any population growth prediction as a factor influencing on *ASMR*. Figure 14 shows the prediction of average smoothed rates of mortality declines for European countries. Average value is taken among following countries: Austria, Belgium, Denmark, Finland, France, Germany, Greece, Iceland, Ireland, Italy, Luxembourg, Netherlands, Norway, Portugal, Spain, Sweden, Switzerland, and United Kingdom. Figure 14 illustrates that number of deaths declines more in young ages and and at ages near 60 or 70 for males and females respectively, whereas in the middle ages and around 100 mortality doesn't change so much.

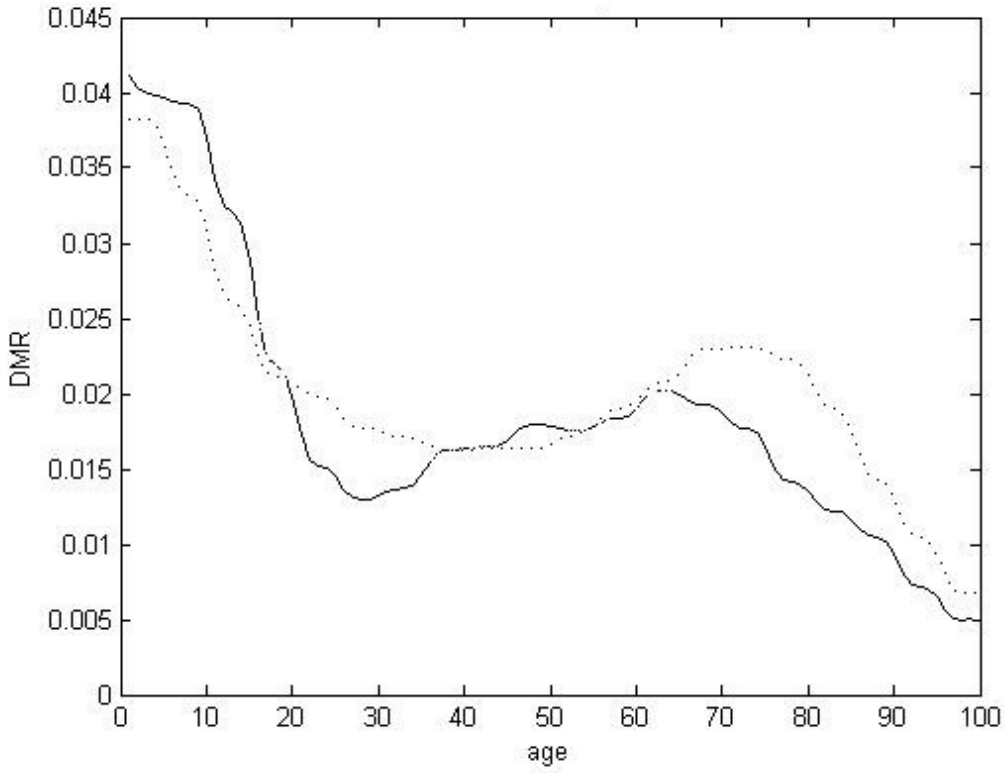

Figure 14. Average decline mortality rates for males (solid) and females (dashed) in Europe in 2004.

Infant mortality rate is the number of deaths of infants under 1 year of age registered in a given year per 1,000 live births registered in the same year. It can be calculated as:

$$IMR = \frac{D_0}{B_0} 1000‰,$$

where D_0 is the total number of infants' deaths during the year, and B_0 is the total number of live births. The number of this measure for Finland in 2004 [24] is

$$IMR = \frac{57\,758}{16\,089} 1000‰ = 3.59‰.$$

Infant mortality rate can be represented on Lexis diagram as a rectangle ABCD in Figure 15.

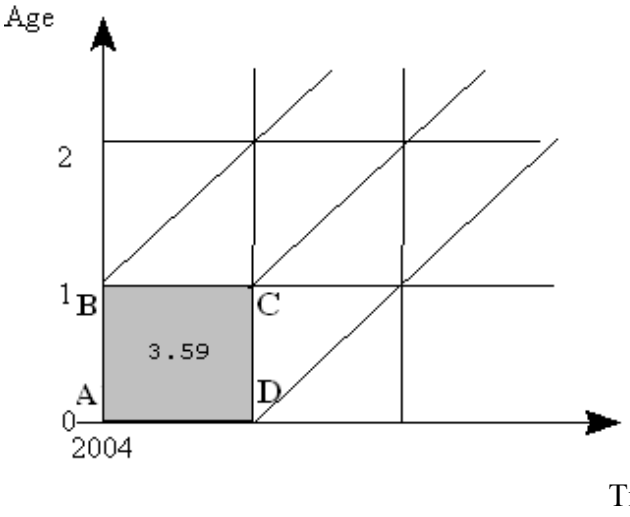

Figure 15. Infant mortality rate.

Infant mortality rate is a very important factor. Mortality in the age below one year heavily exceeds the mortality at other ages except the oldest one. This is very informative and powerful socio-economic factor of the development of the country. Mortality rate for newborns is calculated differently from other ages. This is probably because the number of newborns' deaths is divided by the number of births but not by the average annual number of children. The point is that the average annual number of children at the age below one year difficult to identify, the level of mortality in the beginning and the end of the first life year differs a lot.

Survival probability is a probability of surviving from birth to age x. Let's consider the case when age interval becomes very short so that x is a continuous variable. In that case the total number of years lived during the next n years by those who have reached the age x is:

$$\int_{x}^{x+n} P(a)\,da = P(x)\,\Delta x, \quad \text{when} \quad n \to 0.$$

The instantaneous rate of mortality at a certain age $\mu(x)$ called *force of mortality* and it is calculated like this:

$$\mu(x) = \lim_{\Delta x \to 0} \frac{P(x) - P(x+\Delta x)}{P(x)\,\Delta x} = \frac{-dP(x)}{P(x)\,dx} = \frac{-d \ln P(x)}{dx}.$$

From this formula we can obtain the differential equation, which one can serve as

the definition of the survival probability $P(x)$:

$$\mu(x)\,dx = -d\ln P(x).$$

Let's integrate if

$$P(x) = Const\, e^{-\int_0^x \mu(a)da}.$$

For x=0, we obtain $Const$=$P(0)$, but it is possible to assume that the probability of surviving from birth to age 0 is 1, because *IMR* is considered separately, so $P(0)$=1 and

$$P(x) = e^{-\int_0^x \mu(t)dt}.$$

Let's multiply both sides of equation by $e^{\mu(x)/2}$ and expand the exponentials to the first two terms:

$$P(x)(1+\mu/2) = 1-\mu/2$$

and we obtain so-called *actual estimator* [14]:

$$P(x) = \frac{2-\mu(x)}{2+\mu(x)}.$$

5. Migration

Migration is movement of persons from one country, place or locality to another. Here migrant and migration are different notions, contrary to death and deceased person. The number of migrants does not necessarily equal to the level of migration because one migrant is able to do several migrations during the considered period. The number of migrants is more than or equal to the level of migration.

Data collection about migration is quite difficult. If there is a good system for migration fixation, then detection of the number of migrants is not so clear, and vice versa. Thus, indirect methods based on comparison of two population censuses should be used.

Migration is divided into immigration and emigration. *Net migration* (*NM*) is the difference between immigration and emigration:

$$NM = Immingration - Emigration\,.$$

Net migration measures the influence of space mobility to population size dynamic and it does not depend on the population size unlike other absolute characteristics. Now the main forecast equation **(1)** can be represented in the following way:

$$\begin{aligned} &Population\ size\ at\ year\ t = Population\ size\ at\ year\,(t-1) + \\ &Number\ of\ newborns - Number\ of\ deaths + \\ &Net\ Migration. \end{aligned} \qquad \textbf{(2)}$$

Age-specific net migration rate is a net migration rate limited to a particular age group and for males and females separately. The numerator is the level of net migration in that age group, and the denominator is the number of persons in that age group in the population:

$$ASNMR = \frac{{}_{n}NM_{m}}{{}_{n}P_{m}} \times 1000\,‰\,,$$

where ${}_{n}NM_{m}$ is the level of net migration in the age group [n, m], ${}_{n}P_{m}$ is the number of individuals in the same age group [n, m], where $x \le n < m \le y$. Usually 1 or 5-year groups are used.

We are interested in net migration data, real data of males' and females' net migration in Finland during 2004 are represented in Figure 16.

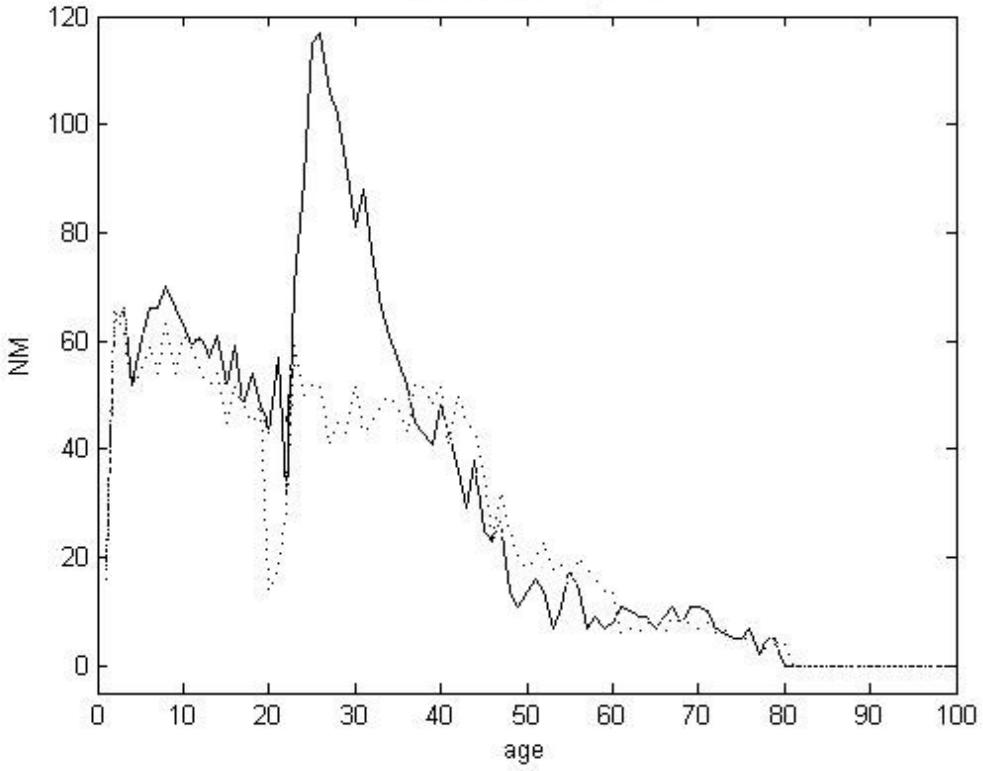

Figure 16. Number of net-migration for males (solid) and females (dashed) in Finland in 2004.

6. Demographic forecast

Demographic forecast is a reasonable scientific assumption about future demographic situation: the number of population, age and sex structure, and the vital rates. It is needed for social and economic plans, education and health programs, for house-building and for retirement insurance, for example. From technical point of view, demographic forecast is a population calculation based on the number of present population, age-sex structure and assumptions about future vital rates.

Demographic forecast is calculating the number of survivors, births and migrations in the considered cohort in each period. Usually period of 1 year is considered. We should take care of three things: the *jump-off population* from which the forecast starts, a set of assumptions about population changes during the period covered by the forecasting, and a method by which the assumptions are applied to the jump-off population. The assumptions about vital rates may be quite simple or very detailed, depending on the level of details required in the final forecast result.

6.1. Linear growth model

A simple model of how population changes over time is *linear growth model*. It was formalized by Leslie in the 1940's [8]. It is based on age-specific survival probabilities and fertility rates, and does not take into account net-migration. The relevant data for each age-group are derived from the the fertility rate and the rate of survival from the previous age-group.

Let's consider a simple toy-example of female population forecast for 5 years. The assumption here is that as females actually give birth, they are more essential to the forecast than males [32]. We suppose that there are 5 age-groups in some hypothetical country with female population distribution as shown in Figure 17. We define population size at forecast year t as a 6-dimensional vector

$$Pop(t)=(Pop(0,t)', Pop(1,t)', \ldots, Pop(5,t)')'.$$

The jump-off population is the population of year t=0. Our goal is to forecast the population from time t to time t+1. We assume that the unit of time is the same as the width of age-group.

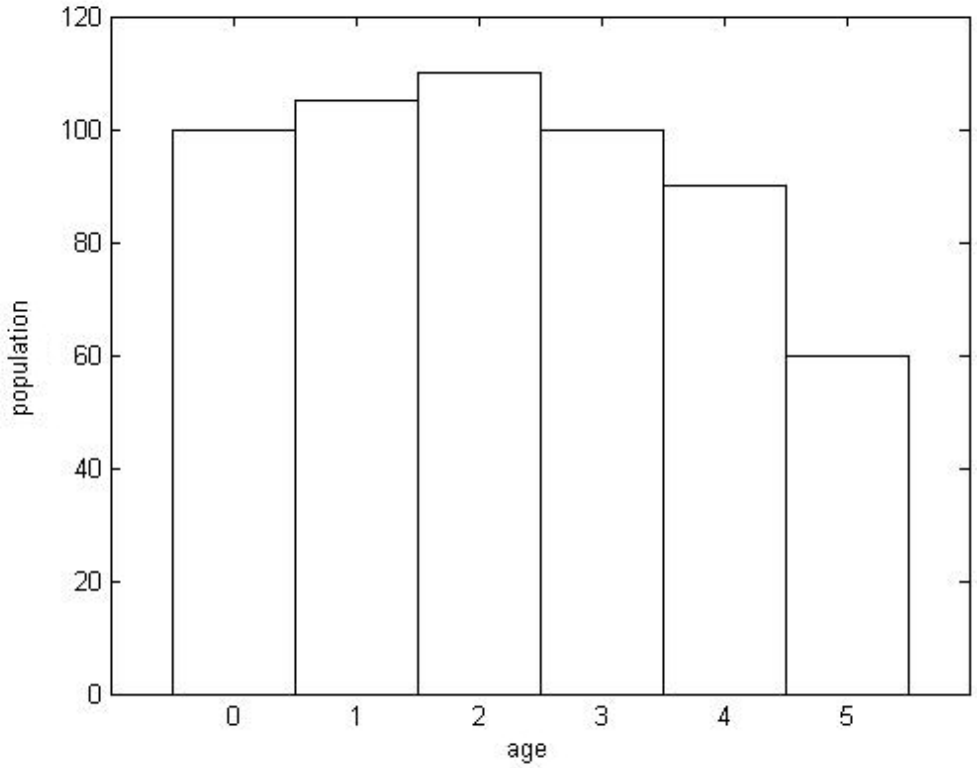

Figure 17. Female population distribution in a hypothetical country.

The persons in age-groups 1, 2, 3, 4 and 5 at time t+1 are the survivors of the previous age-group at time t, we do not consider net migration. That is,

$$\begin{aligned} Pop(1,t+1) &= P(0,t)\,Pop(0,t) \\ Pop(2,t+1) &= P(1,t)\,Pop(1,t) \\ Pop(3,t+1) &= P(2,t)\,Pop(2,t) \\ Pop(4,t+1) &= P(3,t)\,Pop(3,t) \\ Pop(5,t+1) &= P(4,t)\,Pop(4,t), \end{aligned}$$

where $P(x, t)$ is the survival probability from age-group x to the age-group x+1 at the time t.

The new individuals of age-group 0 are newborns and can be estimated with age-specific fertility rates. We assume that the lowest age of childbearing is 2 and the highest age is 4, thus, the number of persons in age-group 0 can be calculated as:

$$P(0,t+1) = F(2,t)\,Pop(2,t) + F(3,t)\,Pop(3,t) + F(4,t)\,Pop(4,t),$$

where age-specific fertility rates are denoted as $F(x)$.

Now we can calculate population size in time t+1 based on the population size in

the previous year t and the values of age-specific fertility rates and survival probabilities. Figure 18 illustrates this process in a graphical way.

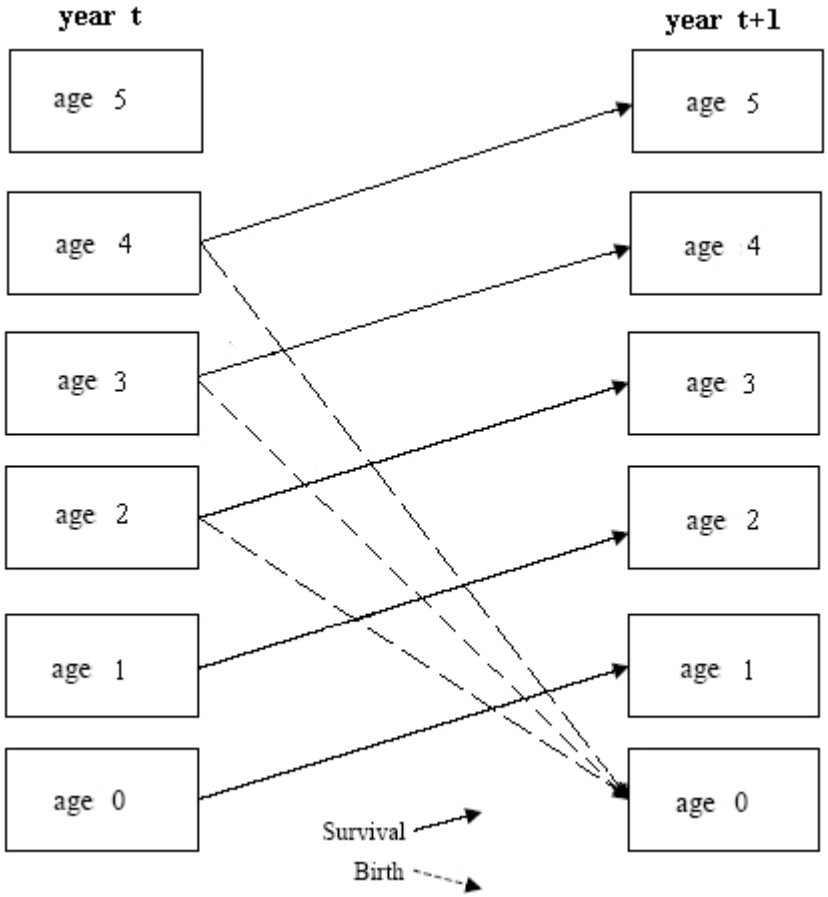

Figure 18. One step of linear growth model without taking migration into account.

The equations can be conveniently written in a matrix form. The number of population is considered as a vector and it is multiplied by a matrix which is empty (contains zeros) except the subdiagonal elements that contain the survival probabilities, and the top row that contains the age-specific fertility rates:

$$\begin{pmatrix} Pop(0,t+1) \\ Pop(1,t+1) \\ Pop(2,t+1) \\ Pop(3,t+1) \\ Pop(4,t+1) \\ Pop(5,t+1) \end{pmatrix} = \begin{pmatrix} 0 & 0 & F(2,t) & F(3,t) & F(4,t) & 0 \\ P(0,t) & 0 & 0 & 0 & 0 & 0 \\ 0 & P(1,t) & 0 & 0 & 0 & 0 \\ 0 & 0 & P(2,t) & 0 & 0 & 0 \\ 0 & 0 & 0 & P(3,t) & 0 & 0 \\ 0 & 0 & 0 & 0 & P(4,t) & 0 \end{pmatrix} \begin{pmatrix} Pop(0,t) \\ Pop(1,t) \\ Pop(2,t) \\ Pop(3,t) \\ Pop(4,t) \\ Pop(5,t) \end{pmatrix}.$$

This can be more compactly as:

$$Pop(t+1) = R(t)\,Pop(t). \qquad \textbf{(3)}$$

From this equation, we can obtain population size in time t=5 recursively form the jump-off population:

$$Pop(5)=R(4)R(3)R(2)R(1)R(0)Pop(0).$$

In more general case this is:

$$Pop(T)=R(T)R(T-1)...R(0)Pop(0). \quad \textbf{(4)}$$

The basic idea of this method is best explained by the means of a simple example. Consider the constant case, when $R(t)=R(t+1)$. It is assumed that the net migration is zero in every age-group and can be ignored entirely. A way of taking migration into account will be introduced later. We need the following data as input :

- Number of forecast years
- Lowest and highest age of childbearing
- Number of age-groups
- Jump-off population
- Survival probabilities
- Age-specific fertility rates

In our example, the number of forecast years is 5, the lowest age of childbearing is 2, the highest age is 4, the number of age-groups is 6 and the distribution of the population size is given in Figure 17. Thus, we need to define survival probabilities and age-specific fertility rates. Figure 19 illustrates our assumptions.

In constant linear growth model, the equation **(4)** is:

$$Pop(T)=R^T Pop(0)$$

or in our case:

$$\begin{pmatrix} 0 & 0 & 1.1 & 1.5 & 0.7 & 0 \\ 0.6 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.8 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.8 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.6 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.5 & 0 \end{pmatrix}^5 \begin{pmatrix} 100 \\ 105 \\ 110 \\ 100 \\ 90 \\ 60 \end{pmatrix} = \begin{pmatrix} 349 \\ 158 \\ 92 \\ 102 \\ 77 \\ 12 \end{pmatrix}.$$

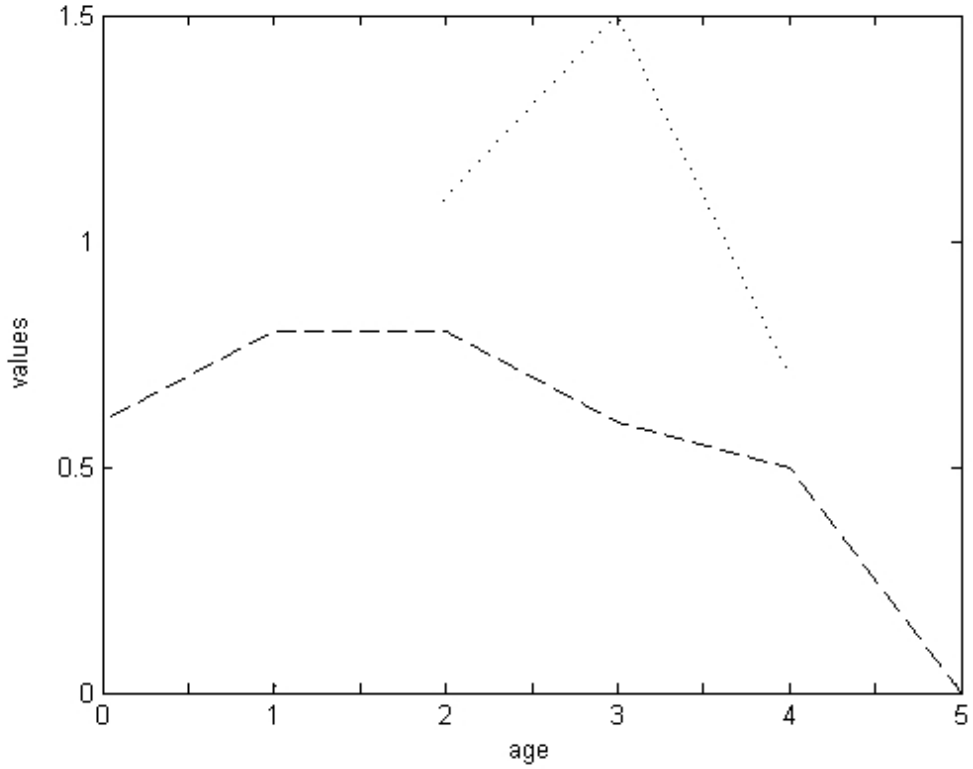

Figure 19. Age-specific fertility rates (dotted line) and survival probabilities (dashed line) for the example.

6.2. Dynamic stochastic population forecast

The considered model is very simple and just illustrates the basic principles. The assumption of zero migration is also not realistic. Migration is extremely difficult to forecast accurately as it is affected greatly by inherently unpredictable events such as war, economic condition, natural catastrophe and changes in administrative restrictions of immigration.

Moreover, age-specific fertility rates and survival probabilities change over time. It should therefore be taken into account, especially when the number of forecast years is large. The simplest way to take migration into consideration is to make assumptions about the net number of migrants by age-group and sex for each forecast year. Then, the number of net-migration in time t is defined as $N(t)=(N(0, t), N(1, t), ..., N(\omega, t))$ and the equation **(3)** can be replaced by

$$Pop(t+1)=R(t)\,Pop(t)+N(t). \qquad \textbf{(5)}$$

The variation of vital rates over the forecast period leads to a dynamic approach, where the elements of matrix R are functions of time. One of the main factor, which influences the population size is human behaviour. We try to predict human behaviour, especially the behaviour of large group of people, which has a low degree of predictability [11]. We consider *stochastic variation of vital rates*, or *dynamic stochastic process*. It means that *forecast error* should be taken into account. Forecast error is a degree of forecast inaccuracy, observed uncertainty, which will be considered in Chapter 9 more detailed. Let us analyse dynamic of age-specific fertility rates, survival probabilities and net-migration.

Consider *ASFRs* according to two characteristics: *TFR* and *MA*. We can consider *ASFR's* as distributions of *TFR* by ages related to the sum of fertility $\Delta_x(t)$, see Chapter 3. Let us call $\Delta_x(t)$ *weights* for simplicity. Thus, the total fertility rate influences on the *density* of *ASFR,* and *MA* influences on the mean value. Figure 20 illustrates how *TFR* affects the linear dimensions the *ASFR*, preserving the outline. Figure 21 shows how *MA* changes the shape of the age-specific fertility rates but conserves the area under the *ASFR* curve. Consequently, it is possible to control *ASFR* curve behaviour with these two parameters.

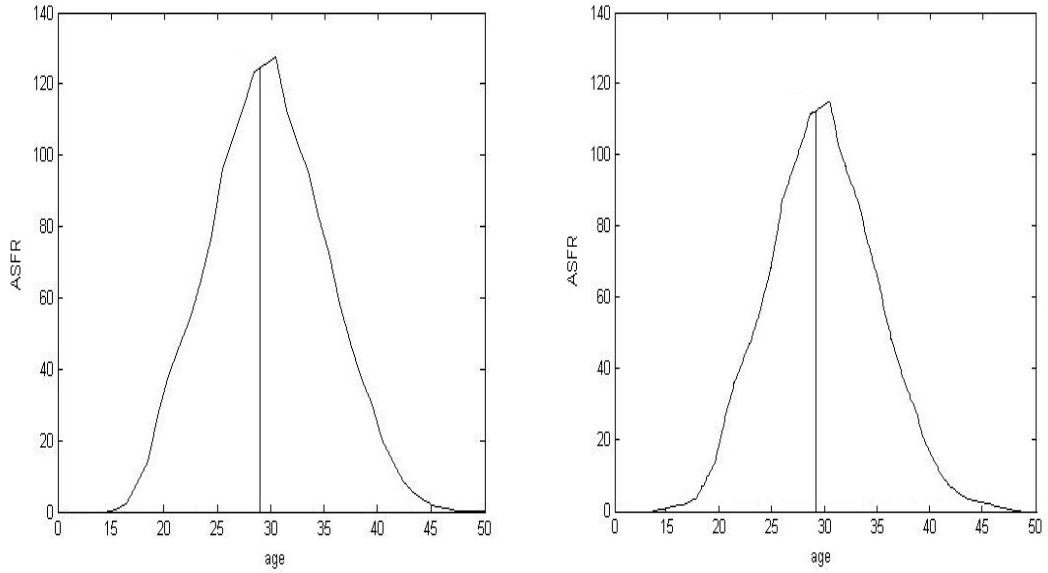

Figure 20. ASFR with the same weights $\Delta_x(t)$ and MA=29.1 but different total fertility rates TFR=1.72 (left) and TFR=1.03 (right).

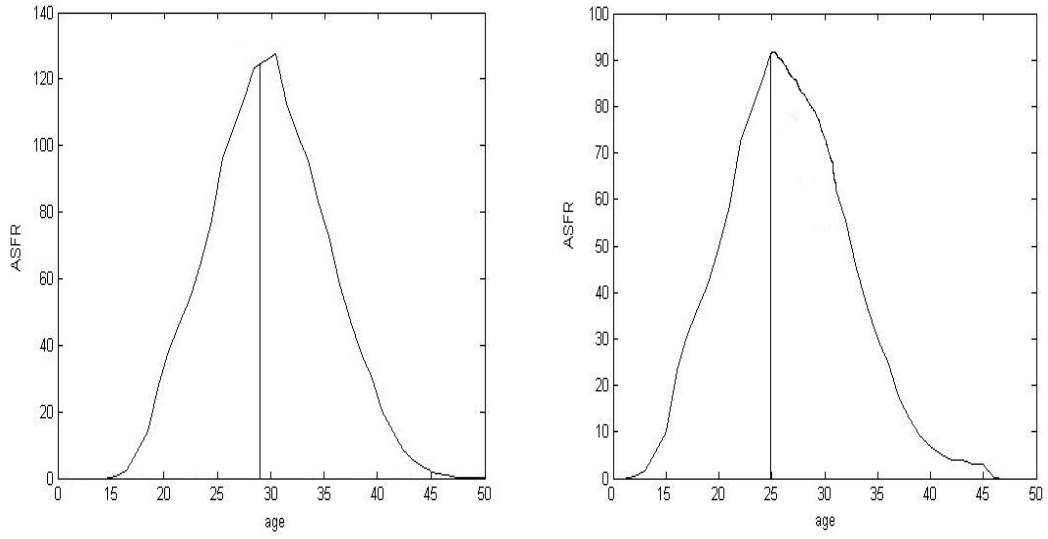

Figure 21. ASFR with the same *TFR,* but different weights $\Delta_x(t)$ and mean age MA=29.1 (left) and MA=24.09 (right).

If we know the initial *ASFR's* in time t=0, and have some guess for the total fertility rate $TFR(T)$ and the mean age $MA(T)$, it is possible to calculate age-specific fertility rates $F(t)$ in time t=1, 2, ..., T, in such a way that the population will reach the guessed $TFR(T)$ and $MA(T)$ in time t=T. We can do it by changing the initial weights $\Delta_x(t)$ linearly to the weights $\Delta_x(T)$, which distribute $TFR(T)$ to $F(t)$ with the guessed $MA(T)$. In other words:

$$\Delta_x(T)=\frac{F(x,T)}{TFR(T)},$$

$$\sum_{x=\alpha}^{\beta}(x+0.5)\Delta_x(T)=MA(T).$$

We consider these assumptions in the next chapter in more details. The last step in the prediction of the age-specific fertility rates is to add the uncertainty into the equation:

$$FF(x,t)=F(x,t)\,e^{error_{fert}(x,t)},$$

where $FF(x, t)$ is the forecast age-specific fertility rate and $error_{fert}(x, t)$ is the forecast error for the fertility rates. It is calculated according to a covariance error prediction model, see Chapter 7.

In our model, the probability of surviving $P(x,\ t)$ equals to the so-called actual estimator of survival, see Chapter 4:

$$P(x,t)=\frac{2-\mu(x,t)}{2+\mu(x,t)},$$

where $\mu(x,\ t)$ is age-specific mortality rate. We assume that *ASMR's* decrease exponentially every forecast year. Thus, if we know the initial values of $\mu(x,\ 0)$ and decline mortality rates for each forecast year, we can estimate $\mu(x,\ t)$ for all t=1,..., T:

$$\mu(x,1)=\mu(x,0)e^{-DMR(x,1)},$$
$$\mu(x,2)=\mu(x,1)e^{-DMR(x,2)}=\mu(x,0)e^{-DMR(x,1)}e^{-DMR(x,2)}=\mu(x,0)e^{-(DMR(x,1)+DMR(x,2))}$$
$$\ldots$$
$$\mu(x,t)=\mu(x,0)e^{-(DMR(x,1)+DMR(x,2)+\ldots DMR(x,t))}.$$

We assume that *DMR's* change linearly. We need to know the initial decline mortality rates $DMR(x,\ 1)$ and ultimate $DMR(x,\ T)$. These values together with age-specific mortality rates in time 0 $\mu(x,\ 0)$ allow us to calculate *ASMR's* for all forecast period. A more careful description will be presented in Chapter 7. With uncertainty forecast age-specific fertility rates look like this:

$$f\,\mu(x,t)=\mu(x,t)e^{(error_{mort}(x,t))},$$

where $error_{mort}(x,\ t)$ is the forecast error for mortality rates calculated according to the covariance error prediction model of Chapter 7.

In the case of net-migration, we assume that it changes linearly over the forecast period. Thus, we need initial and ultimate values of the net-migration number to forecast the level of migration, and then add the forecast error:

$$FN(x,t)=N(x,t)+error_{mirg}(x,t).$$

7. Producing a stochastic forecast of population

In this chapter we consider dynamic stochastic population forecast described by equation **(5)** in more detailed using the extended example from Chapter 6. Forecasting can never completely eliminate risk, it is necessary to consider the uncertainty remaining subsequent to the forecast. This implies that forecasting system should provide also a description of uncertainty as well as a forecast. Hence, our main equation transforms to:

$$Pop(x,t+1)=R(x,t)\,Pop(x,t)+N(x,t)+error(x,t), \qquad \textbf{(6)}$$

where t=1, 2, ..., T is the forecast year, x=0, 1, ... ω is the age-group and $error(t)$ is the forecast error. The starting point of a forecast is that error cannot be avoided, but our goal is to minimize this error. [4]

The goal of forecasting is to estimate the value of $Pop(x, t)$. For this, we need the following the input data. The values in parentheses will be used in the example below:

1. Basic parameters:
    - Number of forecast years (T=5)
    - Highest age of population (ω=5)
    - Lowest and the highest ages of childbearing (α=2 and β=4)
2. Jump-off population ($Pop(x, 0)$)
3. Fertility parameters:
    - Forecast year until which the total fertility rate changes linearly (UT_{TFR}=3)
    - Ultimate total fertility rate ($TFR(UT_{TFR})$=3.8)
    - Forecast year until which the mean age at child-bearing changes linearly (UT_{MA}=3)
    - Ultimate mean age at child-bearing ($MA(UT_{MA})$=3.05)
    - Initial age-specific fertility rates ($F(x, 0)$= (1.1, 1.5, 0.7))
4. Mortality parameters:

- Forecast year until which decline mortality rates change linearly (UT_{DMR}=3)
- Initial decline mortality rates ($DMR(x,\ 0)$=(0.04, 0.037, 0.034, 0.02, 0.017, 0.011))
- Ultimate decline mortality rates ($DMR(x,\ UT_{DMR})$=(0.038, 0.032, 0.029, 0.02, 0.013, 0.008))
- Initial age-specific mortality rates ($\mu(x,\ 0)$=(0.1, 0.4, 0.4, 0.6, 0.8, 0.12))

5. Net-migration parameters:
   - Forecast year until which the number of net-migration changes linearly (UT_{NM}=3)
   - Initial number of net-migration ($NM(x,\ 0)$=(4, 12, 20, 16, 2, 0))
   - Ultimate number of net-migration ($NM(x,\ UT_{NM})$=(2, 10, 16, 10, -2, 0))
6. Uncertainty parameters:
   - *Scales of uncertainty* for fertility, mortality and net-migration

The next step is the definition of matrix $R(x,\ t)$ and the vectors $N(x,\ t)$, $error(x,\ t)$ from equation **(6)**. Next we describe how to forecast the fertility rates, mortality rates, the number of net-migration, and how the uncertainty is handled in dynamic stochastic population forecast.

7.1. Forecast fertility rates

The number of newborns is obtained as a product of forecast *ASFR* and average total population for year 2. We take the average total population for more accurate calculation:

$$Pop(0,t)=FF(x,t)\frac{Pop(x,t-1)+Pop(x,t)}{2},$$

where forecast age specific fertility rates take into account the forecast error $error_{fert}(x,t)$ in the following way:

$$FF(x,t)=F(x,t)e^{error_{fert}(x,t)}.$$

The age-specific fertility rates are distributions of total fertility rate according to the weights $\Delta_x(t)$:

$$F(x,t)=TFR(t)\Delta_x(t).$$

Let's consider the total fertility rates calculation. We assume that it changes linearly. Figure 22 illustrates the total fertility rate over the forecast years. Initial total fertility rate equals to the sum of all initial age-specific fertility rates:

$$TFR(0)=\sum_{x=\alpha}^{\beta} F(x,0)=1.1+1.5+0.7=3.3$$

then for t=1, ..,UT_{TFR}=1, 2, 3:

$$TFR(t)=\frac{((UT_{TFR}-t)TFR(0)+t\,TFR(UT_{TFR}))}{UT_{TFR}}=\frac{3.3(3-t)+3.8t}{3}.$$

If $UT_{TFR} < T$, like in our example, then for $t = UT_{TFR}+1, UT_{TFR}+2, \ldots T$:

$$TFR(t)\equiv TFR(UT_{TFR}).$$

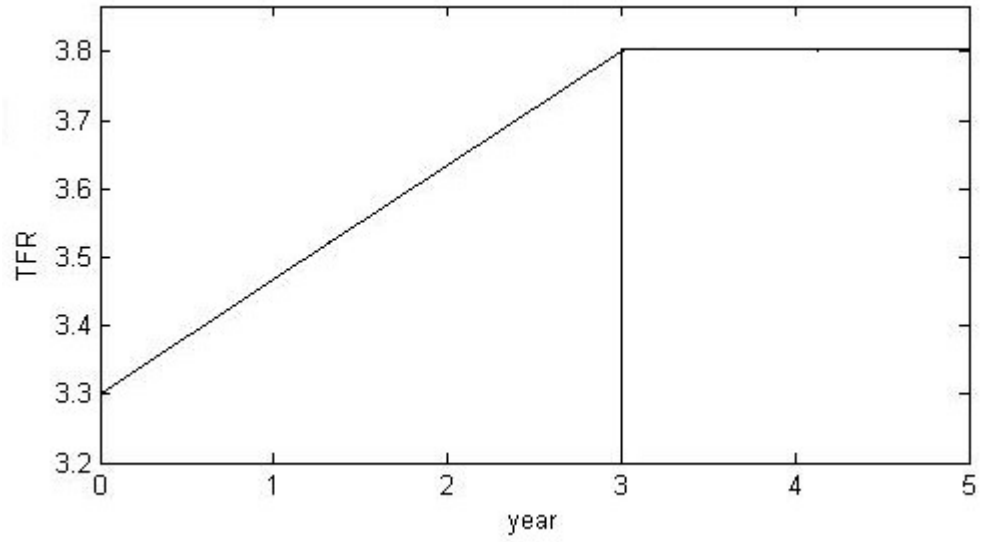

Figure 22. Chang of total fertility rate over the forecast years.

Furthermore, we need to calculate age-specific fertility rates in such a way that the mean age of childbearing will equal to its ultimate value $MA(UT_{MA})$ during the year $t=UT_{MA.}$

In Chapter 3, we presented equation for mean age of child-bearing *MA*. In new notification, initial *MA* can be represented as:

$$MA(0)=\sum_{x=\alpha}^{\beta}(x+0.5)_1\Delta_x=\frac{\sum_{x=\alpha}^{\beta}xF(x,0)}{\sum_{x=\alpha}^{\beta}F(x,0)}+0.5=3.4$$

and we should find $F(x, UT_{MA})$ by solving the following equation for the ultimate *MA* relative to the age-specific fertility rates:

$$MA(UT_{MA})=3.05=\frac{\sum_{x=\alpha}^{\beta}xF(x,UT_{MA})}{\sum_{x=\alpha}^{\beta}F(x,UT_{MA})}+0.5.$$

We take the desired value in the following form:

$$F(x,UT_{MA})=\mathrm{e}^{\gamma(x-MA(0))}F(x,0)=[e^{(-1.4\gamma)}1.1,e^{(-0.4\gamma)}1.5,e^{(0.6\gamma)}0.7],$$

and substitute it in the previous equation. It is possible then to find the parameter γ using Newton's method [29] from the equation:

$$h(\gamma)=\frac{\sum_{x=\alpha}^{\beta}xF(x,UT_{MA})}{\sum_{x=\alpha}^{\beta}F(x,UT_{MA})}-(MA(UT_{MA})-0.5)=$$

$$\frac{\sum_{x=\alpha}^{\beta}xe^{\gamma(x-MA(0))}F(x,UT_{MA})}{\sum_{x=\alpha}^{\beta}e^{\gamma(x-MA(0))}F(x,UT_{MA})}-(MA(UT_{MA})-0.5)=0.$$

We initialize the value γ to $\gamma_{(0)}=0$, and the next values are calculated according to:

$$\gamma_{(i+1)}=\gamma_{(i)}-\frac{h(\gamma_{(i)})}{h'(\gamma_{(i)})}, i=0,1,2,...$$

The stopping criterion is of the form:

$$\left|\gamma_{(i+1)}-\gamma_{(i)}\right|<\frac{1}{1000}.$$

Here the derivative of $h(\gamma)$ has the following view:

$$\frac{(\sum_{x=\alpha}^{\beta} x(x-MA(0))e^{\gamma(x-MA(0))}F(x,UT_{MA}))(\sum_{x=\alpha}^{\beta} e^{\gamma(x-MA(0))}F(x,UT_{MA}))}{(\sum_{x=\alpha}^{\beta} e^{\gamma(x-MA(0))}F(x,UT_{MA}))^2}-$$

$$\frac{(\sum_{x=\alpha}^{\beta} xe^{\gamma(x-MA(0))}F(x,UT_{MA}))(\sum_{x=\alpha}^{\beta} (x-MA(0))e^{\gamma(x-MA(0))}F(x,UT_{MA}))}{(\sum_{x=\alpha}^{\beta} e^{\gamma(x-MA(0))}F(x,UT_{MA}))^2}.$$

Solution is the values of age-specific fertility rates during the year $t=UT_{MA}$, we calculate weights for *ASFR's* $\Delta(x, t)$ for all years t=1, 2, ... $UT_{MA}-1$, assuming their linear growth. In such a way,

$$\Delta(x,t)=\frac{A(x,t)}{UT_{MA}},$$

where

$$A(x,t)=\left(\frac{(UT_{MA}-t)F(x,0)}{F_0}+\frac{tF(x,UT_{MA})}{F_{MA}}\right)$$

and

$$F_0=\sum_{x=\alpha}^{\beta} F(x,0), F_{MA}=\sum_{x=\alpha}^{\beta} F(x,UT_{MA}).$$

If $UT_{MA}<T$, then for $t=UT_{MA}+1, UT_{MA}+2, ..., T$:

$$\Delta(x,t)=\Delta(x,UT_{MA}).$$

Now we are able to derive age-specific fertility rates for all forecast years:

$$F(x,t)=TFR(t)\Delta(x,t).$$

Figure 23 presents the initial *ASFR*(*x,* 0) and the predicted *ASFR*(*x, T*).

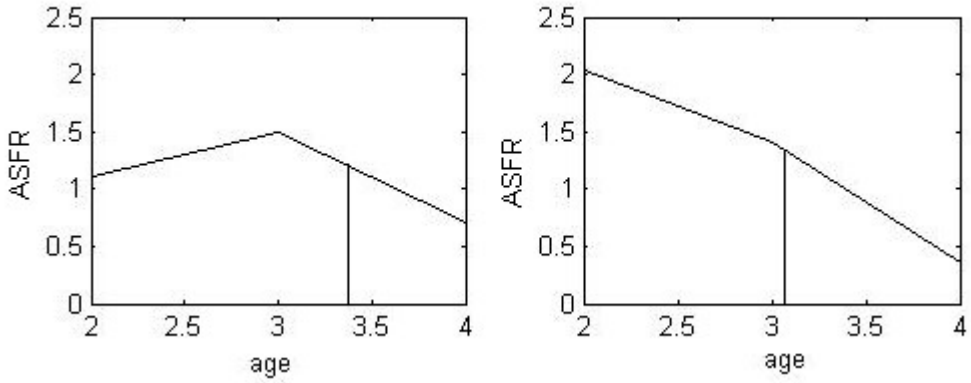

Figure 23. Initial *ASFR* (left) and the forecast *ASFR* (right).

7.2. Forecast mortality rates

Now we consider survival probability *P*(*x, t*). As we have shown in Chapter 4, it equals to

$$P(x,t)=\frac{2-\mu(x,t)}{2+\mu(x,t)},$$

but now we take into account inaccuracy of any forecast, and thus, add the forecast error:

$$f\mu(x,t)=\mu(x,t)e^{error_{mort}(x,t)},$$

where $f\mu(x, t)$ is the forecast age-specific mortality rate, and $error_{mort}(x, t)$ is the forecast error for mortality rates. In such a way, we need to calculate age-specific mortality rates for each forecast year to define the survival probability *P*(*x, t*).

ASMR's are calculated using input data for rates of decline during all forecast period [1*, T*] and initial age-specific mortality rates $\mu(x, 0)$.

We assume that the rates of mortality decline change linearly over the forecast process, thus they are calculated according to the following equation:

$$DMR(x,t)=\frac{(UT_{DMR}-t)\,DMR(x,0)+t\,DMR(x,UT_{DMR})}{UT_{DMR}},$$

where $DMR(x,0)$ and $DMR(x, UT_{DMR})$ are the initial and ultimate rates of mortality decline, and UT_{DMR} is the forecast year until which the rates of decline change linearly. If $UT_{DMR}<T$, a in our example, then for $t=UT_{DMR}+1, UT_{DMR}+2, \ldots T$: $DMT(t)\equiv DMR(UT_{DMR})$. Figure 24 shows linear changes of *DMR* during the forecast period [0,5].

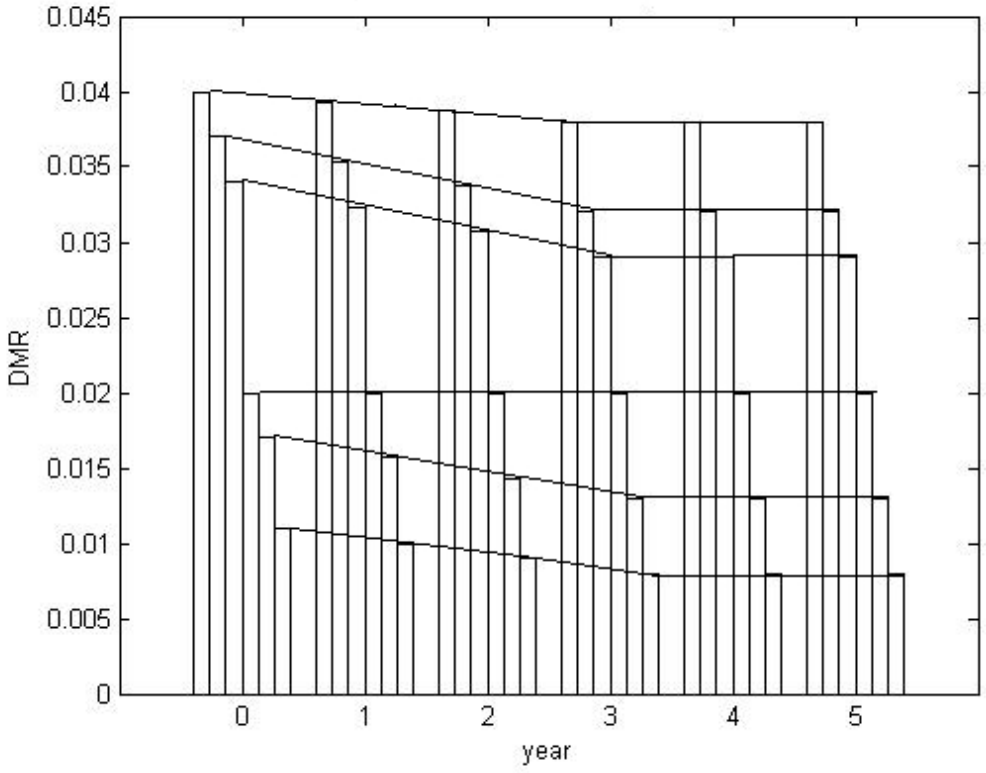

Figure 24. Change of decline mortality rates over the forecast years.

Rates of decline for given age indicates mortality decreasing for population group in that age from year to year and this decrease has exponential nature [27]. Consequently, age-specific mortality rates are calculated as:

$$\mu(x,t)=\mu(x,0)\times e^{-(DMR(x,1)+DMR(x,2)+\ldots+DMR(x,t))}.$$

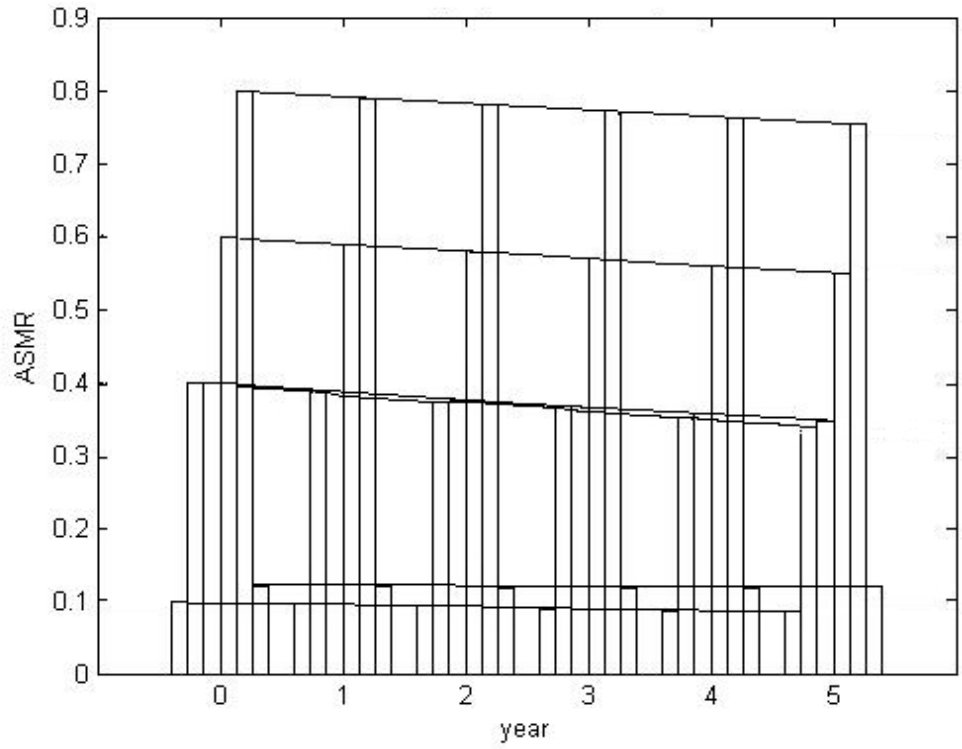

Figure 25. Change of age-specific mortlity rates over the forecast years.

7.3. Forecast the number of net-migration

The next item which we consider is the net-migration:

$$FNM(x,t) = NM(x,t) + error_{migr}(x,t),$$

where $FNT(x, t)$ is the forecast net migration number and and $error_{migr}(x, t)$ is the simulated error for migration.

$NM(x, t)$ is calculated from input parameters of forecast. We assume that the net migration number changes linearly:

$$NM(x,t) = \frac{((UT_{NM} - t)\,NM(x,0) + t\,NM(x,T))}{UT_{NM}},$$

where $NM(x, 0)$ and $NM(x, UT_{NM})$ are the initial and ultimate net-migration number, and UT_{NM} is the forecast year until which net-migration numbers changes linearly. If $UT_{NM} < T$, as in our toy example, then for $t = UT_{NM}+1, UT_{NM}+2, \ldots T$: $NM(t) \equiv NM(UT_{NM})$. Figure 26 shows linear changes of NM during the forecast period [0, 5] and $error_{migr}$ is calculated according to the covariance error prediction model.

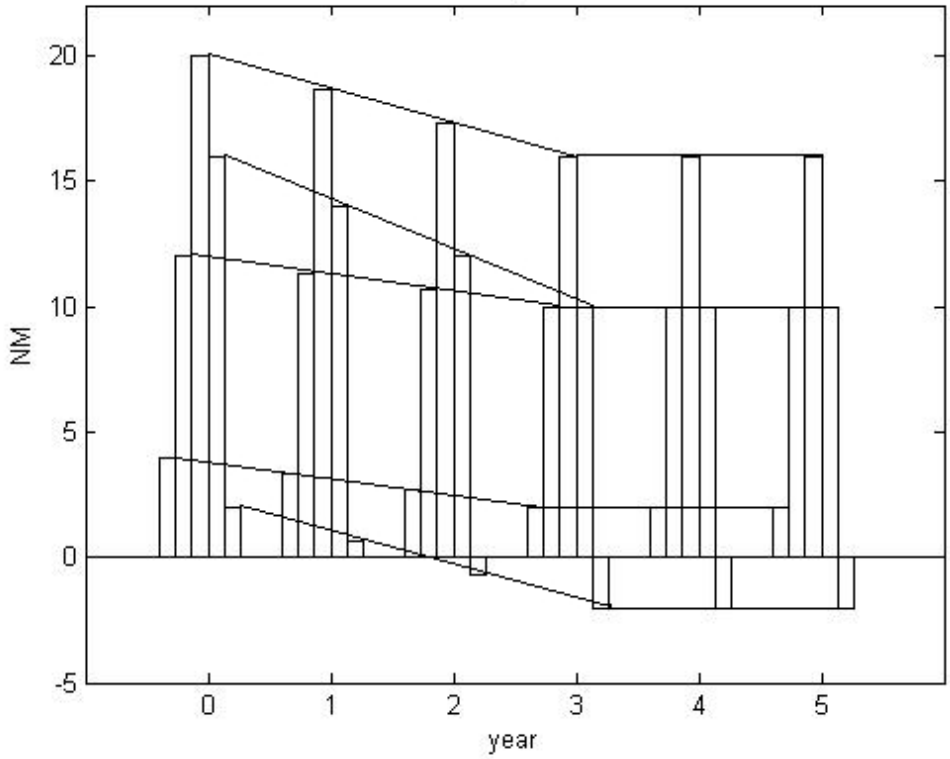

Figure 26. Change of net-migration number over the forecast years.

Now we have all needed information for the construction of matrix $R(x, t)$ except the uncertainty. After consideration of errors we will be able to do the forecast. Next chapter is dedicated to investigation of the uncertainty.

7.4. Uncertainty

Population forecast is a projection in which the assumptions are considered to yield a realistic picture of the probable future development of a population. The forecast is unconditional, i.e. based on current scientific insights, a forecaster gives best guess what the future population will be. Later there can be other opinions, when more information will be available, but at the present the forecast reflects what is currently considered as a plausible future.

Demographic forecasters do in fact give statements about future demographic developments. The statements are not a prior correct, and even plausible. It is a conjecture, a guess or a strong belief based on some calculations. The statements have been tested successfully several times. They are hypotheses, an attempt to explain observed phenomena, and future testing is necessary.

Uncertainty may arise from several reasons [11]:

1. Model misspecification
2. Errors in parameter estimation
3. Errors in expert judgement
4. Random variation

In the first case, the assumed parametric model is only approximately correct. The expected value for variable for which a forecast is made is not correct. There is an error in distribution, so it is impossible to do calculation close to real situation.

In the second case, even if the assumed parametric model would be the correct one, its parameter estimates will be subject to error. The expected values are known, for example, a total number of births in some year, but not time at which an individual women gives birth to a child, which is a random variable. The probability distribution of this random variable is the same for all women involved. When the variance of the distribution is large, the actual number of births will most probably differ much from the expected value. This expected value is known and the uncertainty is expressed by the variance of the distribution.

In the third case, an outside observer may disagree with judgements or prior beliefs about the parameters of the model, or the weights of forecasting. Different forecasters can make different forecasts, because of dissimilar views on the model. Human behaviour cannot be explained, the individuals have a variety of possible actions, and this makes processes describing human behaviour unpredictable. For example, women have a variety of possible actions, which is expressed by the fact that different women have different distribution, or the same distribution, but with different expected values or other parameters. It is logically impossible to infer the probability distribution of the time of childbearing from one single event. In practice, all types of uncertainties will be encountered in a given situation.

In the last case, uncertainty would be left unexplained even if the parameters of the process could be specified without the errors. Since any mathematical model is only an approximation one would expect there to be residual error.

Alho and Spencer proposed the following model for prediction errors [4]. Let $X(j, t)$ be *error processes*, where j=1, 2, ... is the age and $t>0$ is the forecast year. It is assumed that the processes are of the form:

$$X(j,t)=\epsilon(j,1)+\epsilon(j,2)+...+\epsilon(j,t), \qquad \textbf{(7)}$$

where the error increments are of the form:

$$\epsilon(j,t)=S(j,t)(\eta_j+\delta(j,t)).$$

where $S(j, t)$ are *scales of uncertainty*. It is assumed that for each age j, the variables $\delta(j, t)$ are independent over time t=1, 2, ... Variables $\{\delta(j, t)| j=1, 2, ..., \omega; t=1, 2, ...\}$ are also independent from the variables $\{\eta_j| j=1, 2, ... \omega\}$. Furthermore, it is assumed that

$$\eta_j \sim N(0,k_j), \delta(j,t) \sim N(0,1-k_j),$$

where $0<k_j<1$ are known. The terms η_j and $\delta(j, t)$ are calculated as in the $AR(1)$ *model* [1]. Since the error increments are scaled by $S(j, t)$, this model is called a scaled model for error. Inasmuch as the variables $\delta(j, t)$ are independent over time,

$$k_j=Corr(\epsilon(j,t),\epsilon(j,t+h))$$

for all $h\neq0$. Therefore, k_j can be interpreted as a constant correlation between the error increments. It shows the fact that the forecast errors of vital rates in close ages have tendency to be similar, but in distant ages they differ. Here the parameters $S(j, t)$ and k_j are input data for the forecast. $S(j, t)$ allows to scale the value of error, and k_j allows to control the correlation between error increments, when $k_j=0$ error increments are independent and when $k_j=1$ they are perfectly correlated.

This error model allows to produce several alternative predictions of the number of population. Figures 27-28 illustrate sets of forecast population number for our example based on the proposed error model with different values of scale.

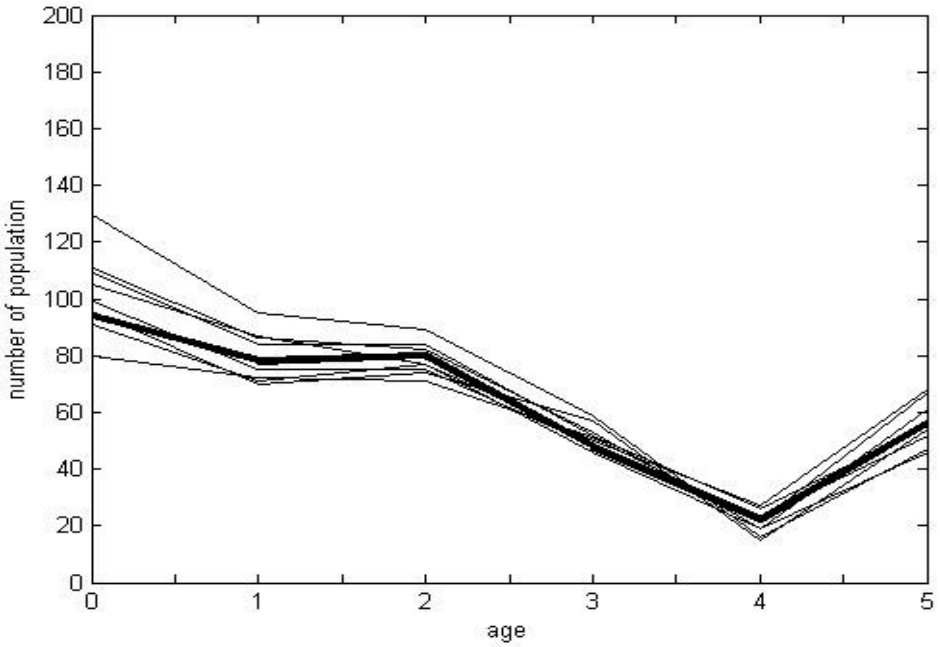

Figure 27. A set of predicted population numbers based on the proposed error model with $S(t, x)_{fer}$=0.06, $S(t, x)_{mort}$=0.033 and $S(t, x)_{mig}$=0.84. Bold curve represents the forecast population number without uncertainty.

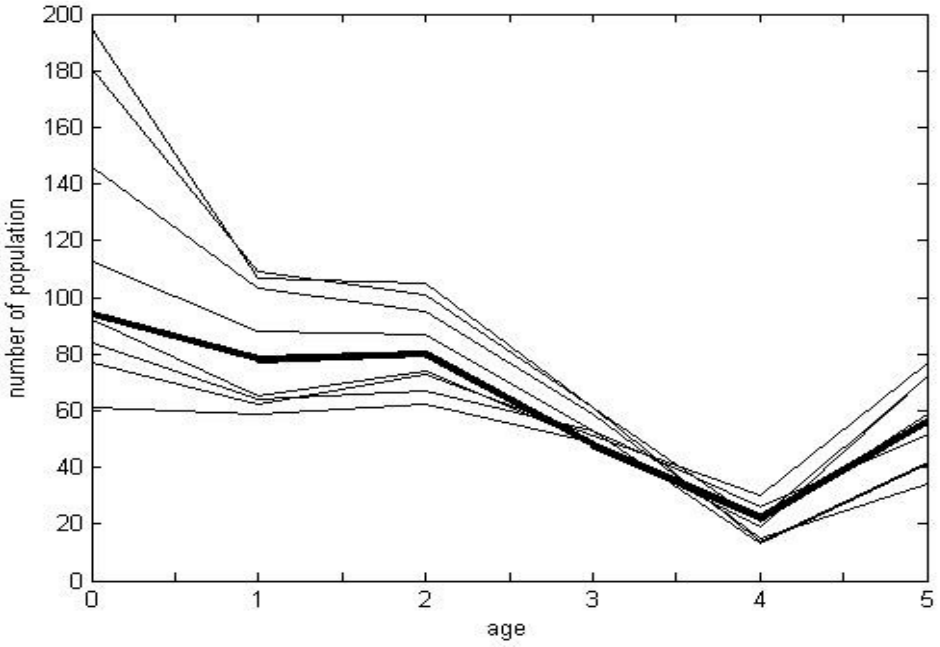

Figure 28. A set of predicted population numbers based on the proposed error model with $S(t, x)_{fer}$=0.12, $S(t, x)_{mort}$=0.066 and $S(t, x)_{migr}$=1.7. Bold curve represents the forecast population number without uncertainty.

8. Computer simulation implementation

Alho and Spencer proposed that population forecasts should be implemented in a computerized database form, where a database can be defined as a collection of data files and computer programs that are capable of storing, updating, and extracting data from the files. An important aspect of the database concept is the possibility to obtain answers in real time [5].

Construction of a forecast database is possible using *simulation techniques*, for example. A simulation technique is an experiment run as a model of reality. The simulations in this paper are computer simulations, they are run on a computer using mathematical models. Simulation techniques are based on the assumption that point forecasts are available for the relevant vital rates, and the user is able to characterize the expected uncertainty of the forecast. Thus, producing new values for the number of population is the result of the simulation.

There is an implementation of a simulation based database forecast in a computer program *Program for Error Propagation* (PEP) [3]. The program produces different sets of results according to a set of assumptions introduced into the system. PEP is based on the described algorithm to forecast population number by sex and age. It is intended for users who use demographic forecasts for planning or scientific purposes, for example demographers, statisticians, economists and actuaries.

PEP produces files with simulated population counts by age and sex for user-specified forecast years; such counts are called *sample paths*. Output files can be read into a statistical program (e.g., *Minitab*) or a spreadsheet program for graphical or statistical description. At this stage, summary information concerning age-groups or the total population can be obtained. Figure 29 illustrates the PEP interface for point forecast input settings.

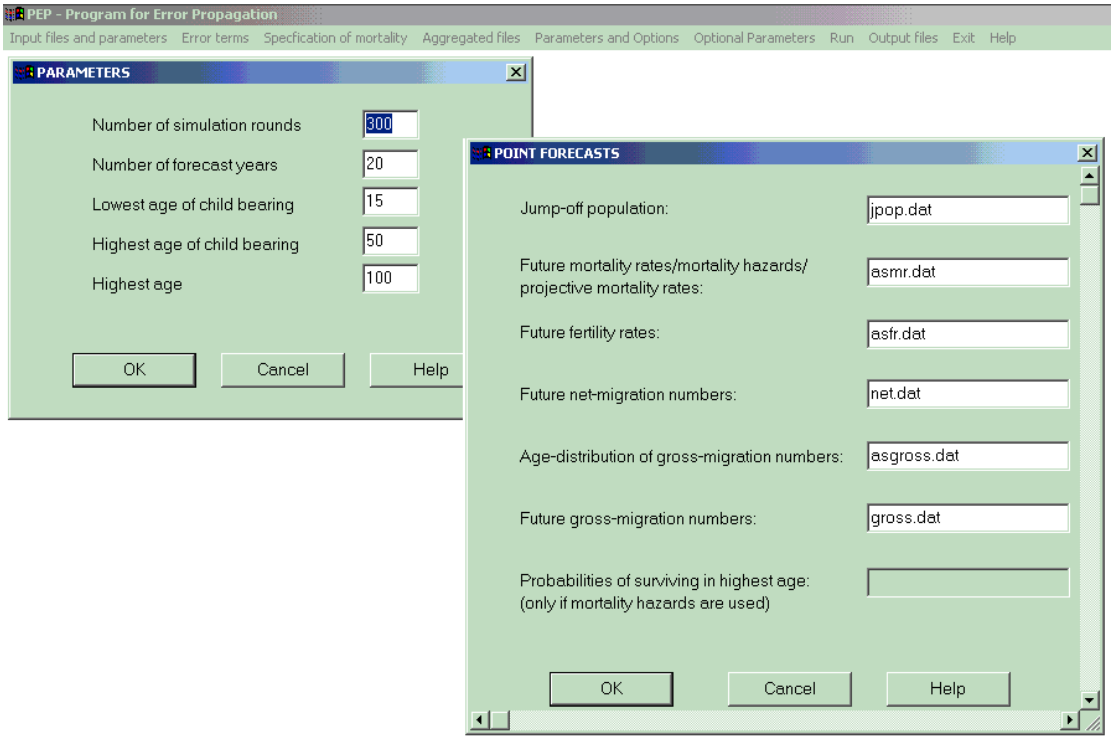

Figure 29. PEP parameters and point forecasts data import.

The main parts of the simulation techniques are the following:

- Reading the point forecast and its preprocessing, see Figure 30
- Generation of the random variables that serve as building blocks in the forecast error simulation
- Calculation of future sample paths of the vital rates based on the point forecast and simulated forecast errors
- Calculation of future population counts using the future sample paths of the vital rates and a linear growth model, see Figure 31. Linear growth model is used for building the matrix $R(x,t)$ for any given t. The vector $N(x,t)$ is calculated as a linear approximation between the initial and ultimate values
- Storing the simulated sample paths for population by age and sex

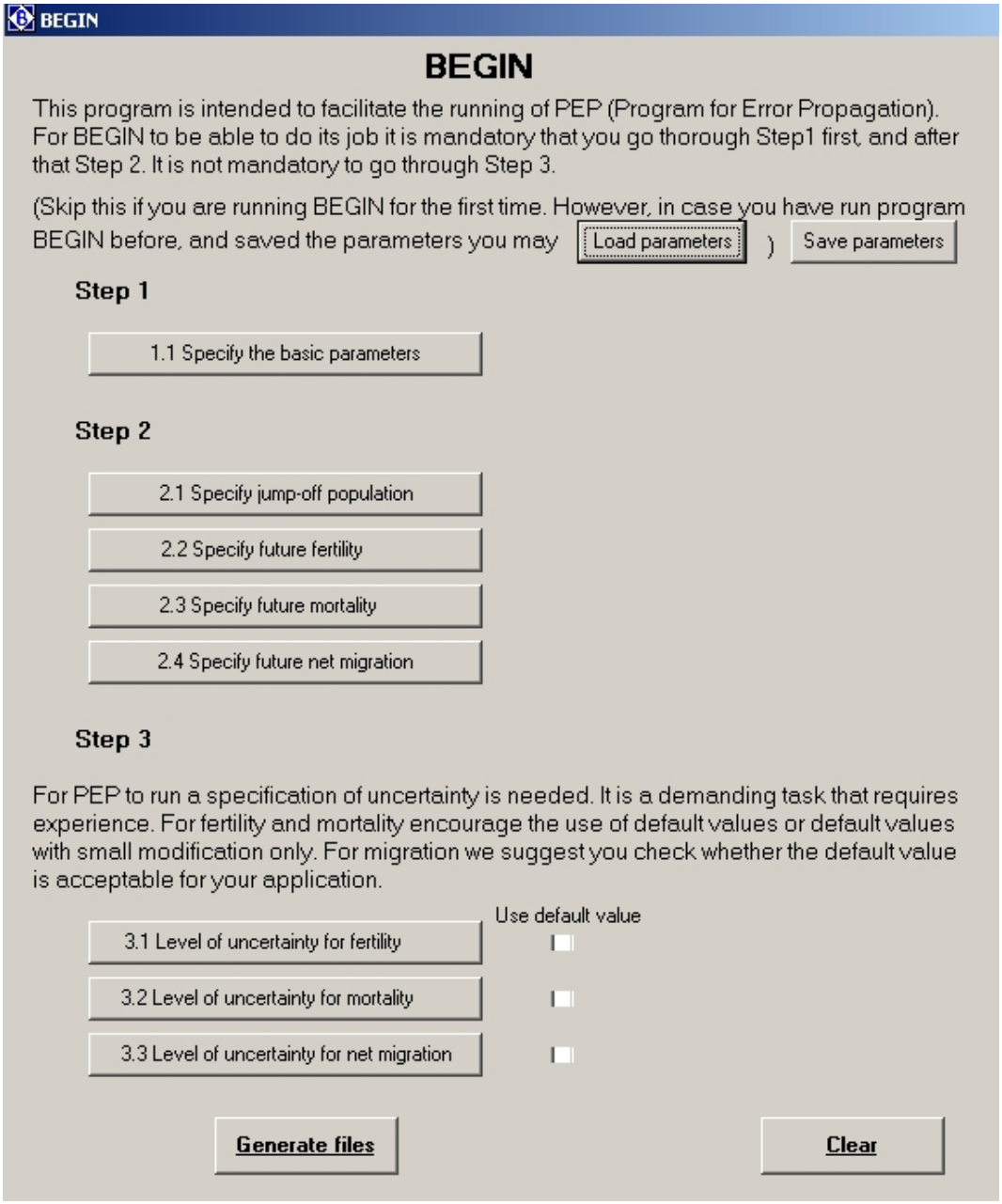

Figure 30. Program interface for reading the point forecast.

Simulation techniques can be considered as generation of sample paths of the forecast error processes for the vital rates, and combining them pathwise via the linear growth model into sample paths to predict the number of population. Having a sufficiently large number of paths available allows us to estimate the underlying predictive distribution with high degree of accuracy [5].

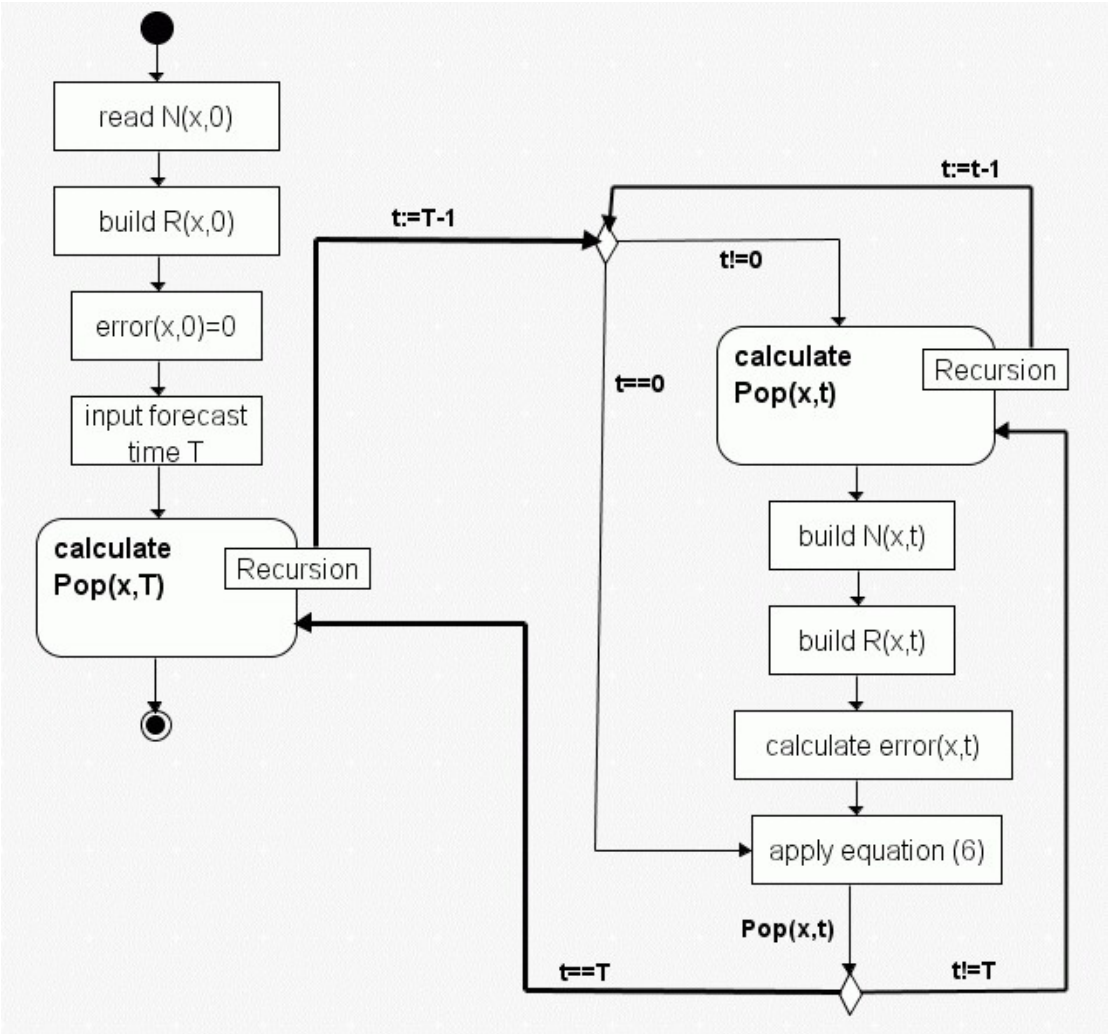

Figure 31. Calculation routine of future population counts.

We recalculate the data of the number of age- and sex-specific group every forecast year based on the data from the previous age group. Thus, we follow every particular cohort during the forecast years according to the forecast vital rates. From the first forecast year t=1, the number of population in every age-specific group is decreased according to the forecast *ASMR*. Thus, we subtract the number of the deaths and migrations (it can be positive or negative value) and persons that survive are moved to the next age-specific group of the population. The number of newborns is calculated according to the fertility rate, infant mortality rate and net migration. We repeat the process for each forecast year, and the result will be the total, age-, and sex-specific number of population for every year of interest. This is illustrated in Table 1, where $P(x, t)$ is the number of population in age x during the forecast year t.

Table 1. Illustration of the simulation technique.

forecast year					
age	1	2	3	...	T
0	$P(0,1)$	$P(0,2)$	$P(0,3)$	...	$P(0,T)$
1	$P(1,1)$	$P(1,2)$	$P(1,3)$	...	$P(1,T)$
2	$P(2,1)$	$P(2,2)$	$P(2,3)$	...	$P(2,T)$
...	...	...	...	...	...
ω	$P(\omega,1)$	$P(\omega,2)$	$P(\omega,3)$	...	$P(\omega,100)$

The program outputs a set of forecasts, where each forecast holds a unique stochastic error value. Each simulation round provides exactly one forecast. After a reasonably large set of forecasts has been generated, it can serve as an input for the various statistical tools to determine the population trends and side flows for any given forecast interval. Alternatively, user can select the one with the maximum degree of certainty.

As an idea for future development of PEP, it would be possible to develop a pluggable architecture for the seamless integration of various statistical post-PEP processing tools.

9. Experiment with Finnish population (for 2004-2054)

Let us construct a sample forecast according to the model discussed earlier using the software. We choose 50 years forecast for 2004-2054. We use two programs: PEP and BEGIN [6]. BEGIN produces the input files in a correct way for the program PEP. After a BEGIN run, we start PEP without any additional information. The output files contain the predictive distribution of the future population by age and sex for each forecast year. Thus, all necessary input data can be specified via BEGIN.

Basic parameters. We set the number of forecast years equals to 50 (2004-2054). The other parameters are set as follow:

- Number of simulation rounds 1500
- Lowest age of child-bearing 15
- Highest age of child-bearing 49
- Highest age of population 100
- Seed for random number generation 1

Jump-off population. The jump-off population is the Finnish population at the end of year 2004 as the number of people at age 0, 1, 2, ..., 99, 100+, females and males separated. This population is the legally resident population as enumerated in the Finnish population register. We assume that the accurate is nearly 100%. In countries with less accurate registration systems, the uncertainty of the jump-off value can also be taken into account. Figure 32 illustrates the starting population for our forecast distributed by age and sex.

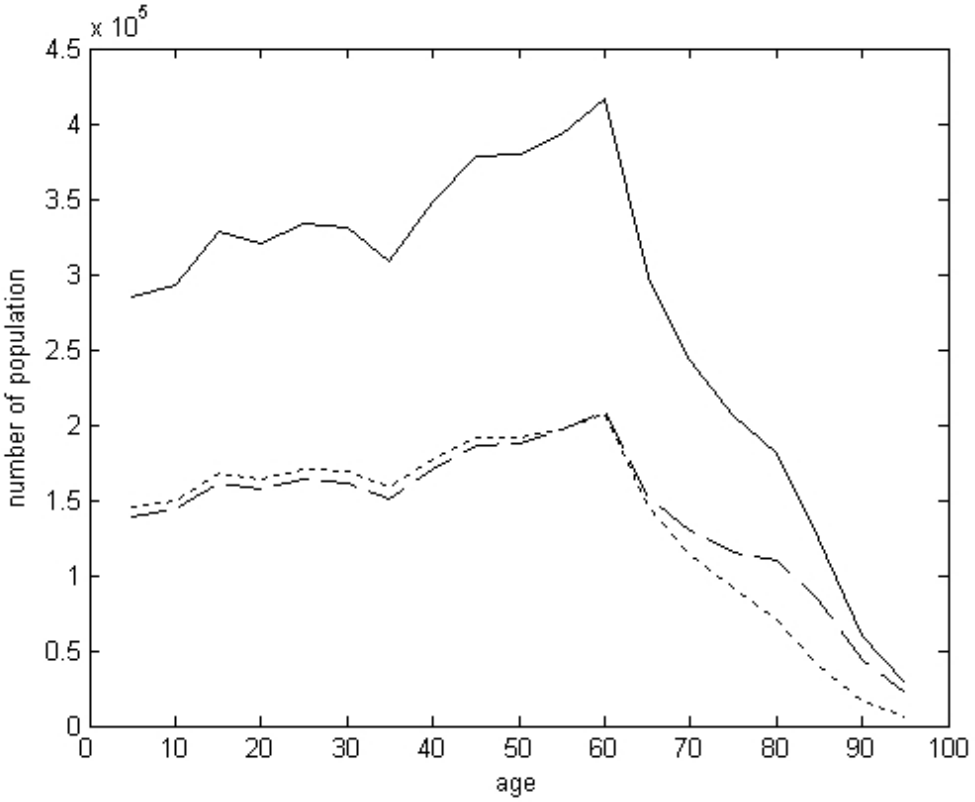

Figure 32. The number of total population (solid line), female population (dotted line) and male population (dashed line) in Finland in 2004.

Fertility. There are several parameters needed for fertility forecasting:

- Initial fertility rates
- Ultimate total fertility rate
- Ultimate mean age of childbearing
- Forecast years until which *TFR* and *MA* change linearly

Figure 10 in Chapter 3 illustrates the age-specific fertility rates in Finland for the starting year 2004. In their forecast, United Nations [25] assumes that the fertility rate in Finland over the first 5 or 10 years of the projection period will follow the recently observed trend 1.73. After this transition period, fertility is assumed to increase linearly at a rate of 0.07 children per woman. Thus, ultimate *TFR*=1.85. Current mean age of child bearing is 29.6 but is expected to change to about 32 by 2050. We assume that the mean childbearing age will also move a little bit, ultimately to *MA*=31. We also assume that *TFR* and *MA* change linearly during the whole forecast period.

Mortality. For mortality forecasting we need to specify:

- Mortality probability values for newborns and projective rates for people from age one to the highest age
- Initial and ultimate values for decline rates
- Forecast years until which the rates of decline change linearly

Figure 12 in Chapter 4 shows mortality probability values for newborns and projective rates for people at age from one to the highest age for males a females in 2004. The task to set the rates of decline is not trivial. 70 years ago female life expectancy was 57 years, and nowadays it is 81 years [25]. Thus, female life expectancy has increased by 81-57=24 years, which corresponds to *DMR*=0.34 years annually. However, if we observe 50 years interval 1954-2004, the increase is 78-57=21 years, which is *DMR*=0.42. During the latter 30 years the increase is 81-75=6 years, which is *DMR*=0.2.

Alho proposed that age-specific mortality rates continue to decline at the rate they have declined during the past 15 years [2]. During the past 15 years female life expectancy improved only by 81.0–78.6=2.4 years, *DMR*=0.16. This implies an improvement to 88 years by 2050. Figure 14 contains *DMR* for all ages for males and females. We assume that the same tendency continues during the whole forecast period, so that the ultimate values of decline rates coincide with the initial.

Migration. The forecasting of migration differs from that of fertility or mortality because of its nature. We can influence to migration via different social impacts more than to fertility and mortality. Another difference is the data on migration are poor even in a country like Finland that has a well functioning population register. Because of these problems, migration forecasts are typically judgemental, and given in terms of the net number of migrants. Alho assumed that the most likely net number of migrants would remain at the recent level of 4,000 per year [2]. United Nations assumes that the future path of international migration should be set on the basis of past international migration estimates and an assessment of the policy stance of countries with regard to future international migration flows.

To specify the forecast for migration we need:

- The initial and ultimate net migration numbers
- Forecast year until which the number of net migration changes linearly

The initial number of net migration can be found in Figure 16 in Chapter 5. Our prediction for the ultimate net migration in Finland are shown in Figure 33. We assume that the net migration changes linearly during the whole forecast period.

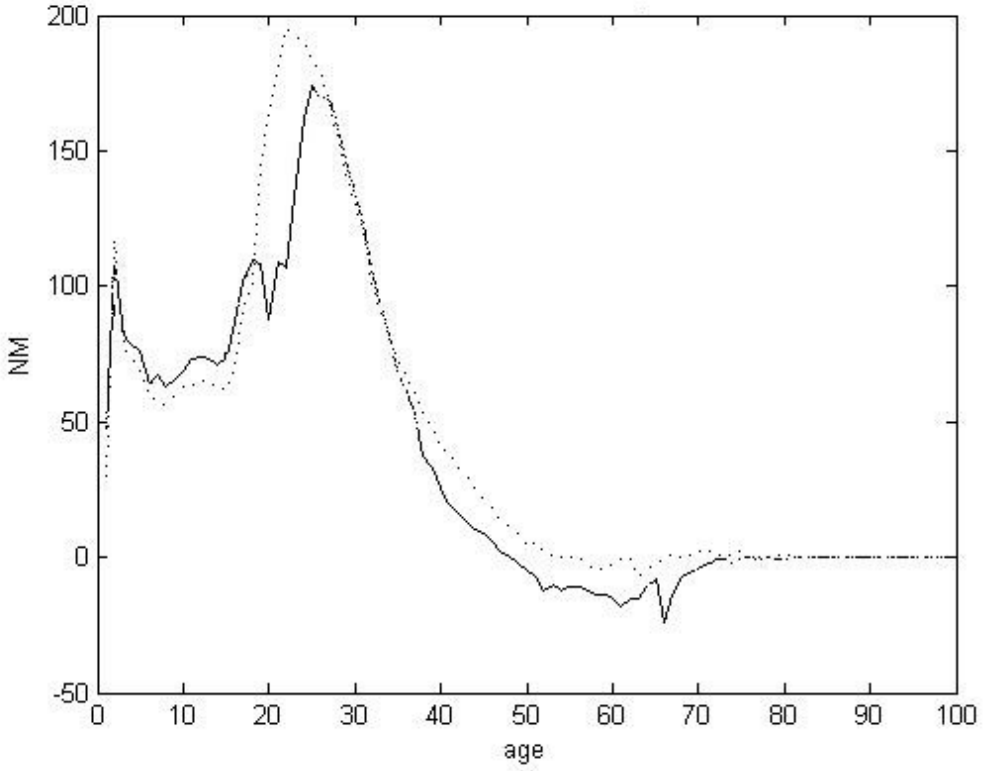

Figure 33. Predicted ultimate number of net-migration for males (solid) and females (dashed) in Finland in 2054.

Uncertainty. Uncertainty is a demanding task that requires certain experience. We follow [9] and set the values as follows:

- Scales of uncertainty $S(j, t)$ for age-specific fertility 0.06
- Scales of uncertainty $S(j, t)$ for age-specific mortality 0.033
- Scales of uncertainty $S(j, t)$ for age-specific net-migration 2

Eventual value for the scales of fertility was obtained from long data series of Denmark, Finland, Iceland, the Netherlands and Sweden. The scales for mortality were estimated from long data series from Austria, Denmark, Finland, France, West

Germany, Italy, the Netherlands, Norway, Sweden, Switzerland and United Kingdom. The estimates were based on the median level of uncertainty in the past, averaged across all countries.

Results. After simulation, PEP produces 50 files, one for each forecast year. Each file contains the predictive distribution of the future population by age and sex. The columns correspond to the age-groups, and the rows to the simulation rounds. The first row is the title row, it contains information about sex and age, the second row contains the future population counts of the first simulation round, the third row the counts of the second simulation round, and so on. In our case, a fragment of a file looks as following:

M0	M1	..	M100	F0	F1	..	F100
35669	32682		17610	32050	33241		23973
15565	16068		1035	15155	14295		1404
....	...	..	...	...	...	..	...
30829	30851		6553	29311	29357		7935

Thus, we have 1500 alternative forecasts of the population by age and sex. We can summarize the results for all ages and both sexes together, and then build predictive distribution. Figure 34 shows a histogram for the predictive distribution of the total population divided into 5-years age-group. Median value equals to 5 553 000, now the total population is P=5 236 611. Thus, low fertility rate will be compensated by the increase of life expectancies and migrations.

The forecast of United Nations is 5 329 000. Figure 35 shows the mean value of United Nations population forecast and the result of the simulations perfomed in this thesis together, divided into 5-year age groups. Their shapes have similar structure but the difference in the number of newborns occures due to the relatively big scale for age-specific fertility. Figure 36 illustrates the minimum, maximum and median values of our popularion forecast in Finland in 2050.

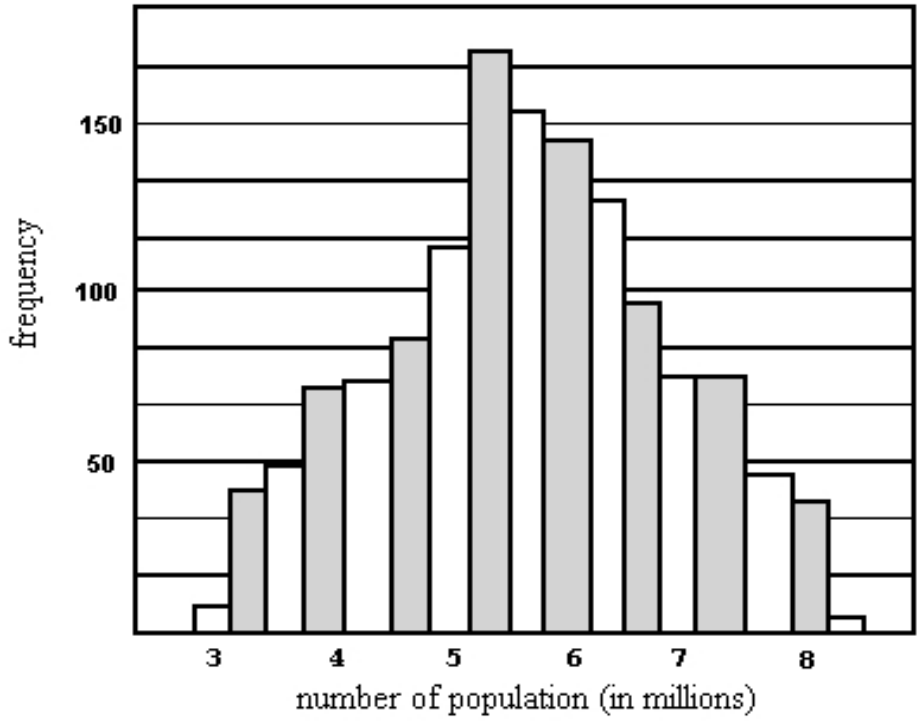

Figure 34. Predictive distribution of total population in millions in Finland in 2054.

Besides the future population counts, PEP produces simulated life expectancies for males and females. The simulated life expectancies are stored into annual files, where y=1, ..., 50 refers to the forecast year. The first column contains the simulated life expectancies for males, and the second column the simulated life expectancies for females. In our case, a fragment of file a looks like the following:

```
85.11     89.02
84.93     89.55
85.04     88.97
85.78     89.82
 ...       ...
85.55     89.60
```

The mean values for male and female life expectancies are 85 and 89. United Nations gives a more pessimistic forecast: 82.1 and 87.1 for the years 2040-2050. It can be explained by underestimation of the mortality decline, which can be observed during the last years in United Nations forecasts.

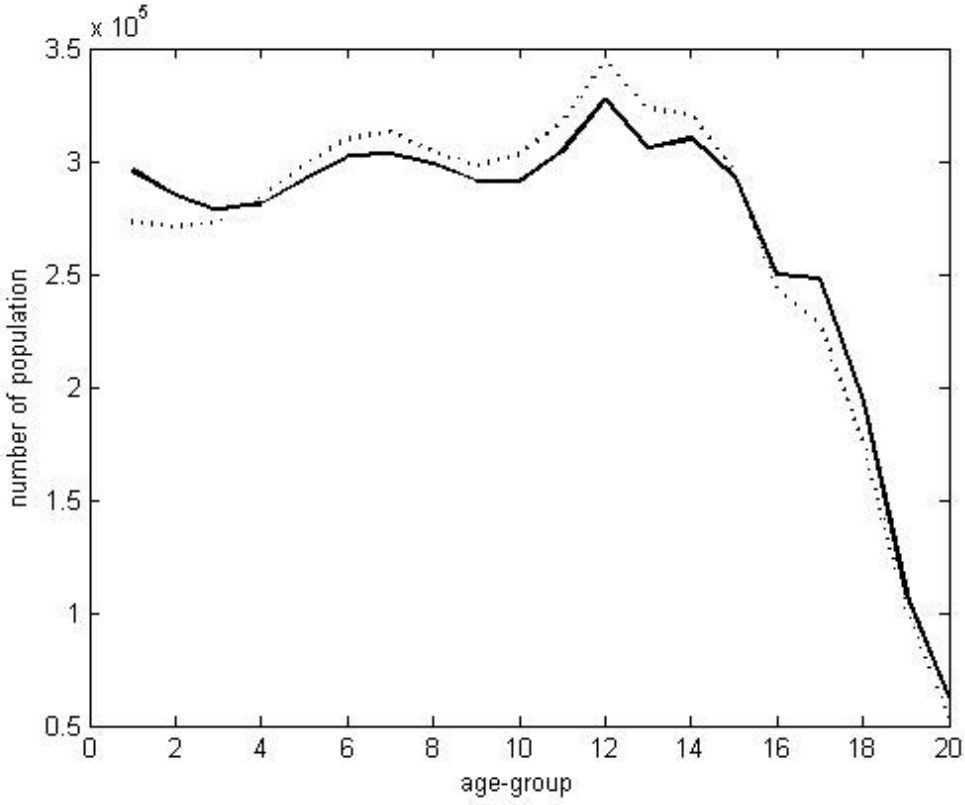

Figure 35. Mean values of United Nations population forecast (dashed line) and the simulated (solid line) in Finland in 2050.

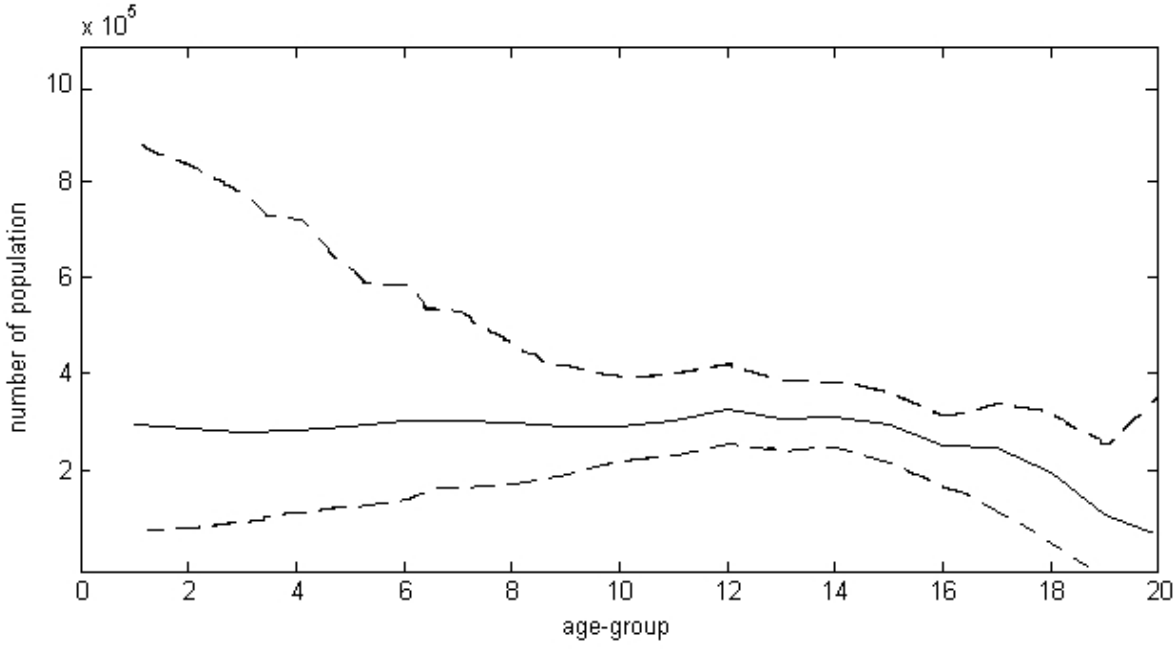

Figure 36. Forecast of total polulation by 5-years age-group: minimum, maximum (dashed lines) and median (solid line) values in Finland in 2050.

10. Conclusion

Stochastic forecast of population is studied and illustrated. A short introduction to the basic demographic concepts is first presented. Lexis diagram is used as main instrument of demographic analysis. Fertility, mortality, migration and their measures are considered in details with real data examples for Finland in 2004. After that, we discuss linear growth model as a basis of our simulations.

Simulation technique is chosen as a forecast model. It is a natural to think about population changing over the forecast period and it is a simple and effective model, that can be easily implemented in a computer. We assume that the total fertility rate, decline mortality rate and net migration change linearly. Age-specific fertility rate changes according to the total fertility rate with defined mean age of childbearing. Age-specific mortality rate decreases exponential. Simulation error is calculated according to the proposed error prediction model.

Experiment with Finnish population for 2004-2054 is given. Calculation of population size is repeated 1500 times and then median values are given as output. It allows to achieve more realistic data, because we have every time new values for the simulation errors calculation. Our result is compared with the United Nation population forecast for Finland in 2050, and they have similar structure.

The current work has shown that the selected topic is of interest to be continued. A future improvement of studied method would be an alternative selection of the output among the simulated population paths.

References.

1. Abraham B. and Ledolter J.: *Statistical Methods for Forecasting.* John Willey & Sons, N.Y., 1983.
2. Alho J.: *A stochastic forecast of the population of Finland.* Statistics Finland, 1998.
3. Alho J.: PEP – program for error propagation. Internet WWW-page, URL: http://joyx.joensuu.fi/~ek/pep/pepstart.htm (13.05.2006).
4. Alho, J. and Spencer, B.: The practical specification of the expected error of population forecasts. *Journal of official statistics*, no 13: 203–225, 1997.
5. Alho, J. and Spencer B.: *Statistical demography and forecasting.* Springer, N.Y., 2005.
6. Yanulevskaya V. and Alho J.: BEGIN tutotrial. Internet WWW-page, URL: http://joyx.joensuu.fi/~ek/pep/BeginTutorial.pdf (13.05.2006)
7. Bongaarts J. and Potter R.: *Fertility, biology, and behaviour. An analysis of the proximate determinants.* Academic Press, N.Y., 1988.
8. Caswell H.: *Matrix population models: construction, analysis and interpretation.* Sinauer Associates, Sunderland, M.A., 2001.
9. Changing population of Europe: uncertain future**.** Final report, Statistics Netherlands, 2005.
10. Gourbin C. and Masuy-Stroobant G.: Are live and stillbirths comparable all over Europe? Legal definitions and vital registration data processing. Institut de Demographie Working Paper, no 170, 1993.
11. Keilman N.: *Uncertainty in National Population Forecasting: Issues, Background, Analyses, Recommendation.* Swets & Zeitlinger, Amsterdam, 1990.
12. Keyfitz N.: *Applied Mathematical Demography.* Springer, N.Y., 1985.
13. Keyfitz N. and Beekman J.: *Demography Through Problems.* Springer, N.Y., 1984.
14. Keyfitz N.: *Introduction to the Mathematics of Population with Revisions.* Addison-Werley, Reading M.A., 1977.

15.Levenbach, H. and James P.: *The modern forecaster: the forecasting process through data analysis.* Van norstrand reinhold company, 1994.

16.Merriam-Webster Inc.: Merriam-Webster on-line dictionary. Internet WWW-page, URL:
http://www.m-w.com/ (12.02.2006).

17. Newell C.: *Methods and Models in Demography.* Chichester: John Wiley & Sons, 1988.

18.Peterson, W. and Renee P.: *Dictionary of Demography: Terms, Concepts and Institutions.* N. Y. : Greenwood Press, 1986.

19.Poikolainen K. and Eskola J.:The effect of health services on mortality: decline in death rates from amenable and non-amenable causes in Finland, 1969-81. *Lancet* I:199-202, 1986.

20.Pollard A., Farhat Y. and Pollard G.: *Demographic Techniques.* Pergamon Press, Sydney, 1981.

21.Shryock H. and Siegel J. : *The methods and materials of demography.* Washington, D.C.: Govt. Printing Press Office, 1980.

22.Statistics Finland. Internet WWW-page, URL:
http://stat.fi/tup/euupe/ (16.09.2006).

23.The Human Mortality Database. Internet WWW-page, URL:
http://www.mortality.org/ (17.04.2006).

24.The world fackbook by Central Intelligence Agency. Internet WWW-page, URL: http://www.cia.gov/cia/publications/factbook (16.09.2006).

25.United Nations. Internet WWW-page, URL:
http://www.un.org (16.09.2006)

26.United Nations. Multilingual Demographic Dictionary, English Section, United Nations Department of Economic and Social Affairs, N.Y., 1958.

27.Vallin J., D'Souza S. and Palloni A.: *Measurement and analysis of mortality: new approaches.* Oxford University Press, N.Y., 1990.

28.Vandeschric C.: *Demographic analyse.* Academia-Bruyland, L'Harmattan, 1995.

29. Wikipedia the free encyclopedia: The Newton's method description. Internet WWW-page, URL: http://en.wikipedia.org/wiki/Newton%27s_method (03.02.2006).
30. World Health Organization. Internet WWW-page, URL: http://www.who.int/en/ (16.03.2006).
31. Дубова Т.: *Статистические Методы Прогнозирования.* Юнити, Москва, 2003.
32. Медков, В.: *Демография.* Москва: ИНФА-М, 2004.